AWARD NUMBER: W81XWH-13-1-0020

TITLE: Health-Terrain: Visualizing Large Scale Health Data

PRINCIPAL INVESTIGATOR: Ph.D. Fang, Shiaofen

CONTRACTING ORGANIZATION: Indiana University, Indianapolis, IN 46202

REPORT DATE: December 2014

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
Dec 2014

2. REPORT TYPE Annual

3. DATES COVERED
7 Mar 2013 - 6 Sep 2014

4. TITLE AND SUBTITLE
Health-Terrain: Visualizing Large Scale Health Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-13-1-0020

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shiaofen Fang, Mathew Palakal, Yuni Xia, Shaun J. Grannis, Jennifer L. Williams

E-Mail: craigjen@regenstrief.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Indiana University
980 Indiana Avenue, RM2232
Indianapolis, IN 46202

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The promise of the benefits of fully integrated electronic health care systems can only be realized if the quality of emerging large medical databases can be characterized and the meaning of the data understood. For this purpose, the effective visualization of large and complex health data for timely decision making is critical. Our long-term goal is to improve the usability of emerging large scale clinical data sets by developing effective and efficient open-source systems for health data analytics and visualization tools for clinicians, healthcare professionals, administrators, and patients. The objective of this application is to develop a prototype system to test the effectiveness of this approach on a large scale health care database that is currently available at Regenstrief Institute. We have reached this objective with the following specific accomplishments:

- Built a relational database as the representation of a health concept space, extracted from the NCD dataset.
- Natural Language Processing techniques were carried out to process 325791 clinical notes to extract new terms including diseases, symptoms, and mental and risky behaviors.
- Data mining techniques were applied to extract associations between terms in the concept space, and to discover new cluster terms.
- Designed and implemented a suite of novel visualization algorithms that allows the users to interactively explore the data based on the user selected terms and filters.
- Designed and implemented a web based graphical user interface for the prototype system.
- Designed and tested an evaluation procedure for health data visualization system.

This visualization framework offers a real time and web-based solution for the effective use of large scale military electronic health record systems by allowing system level integration of the human´s visual capabilities into the overall health data based decision making system. The visual representation of concept space provides a method to compress large, heterogeneous, and historical patient and public health data into a single, intuitive and comprehensive visualization. The new spatiotemporal visualization techniques developed here are novel and particularly suited for large public health datasets that involve geographical and population wide information.

15. SUBJECT TERMS
Information visualization; Visual analytics; Public health data, Notifiable condition detector; Text mining; Data mining.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unclassified	79	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION

The goal of this project is to develop novel visualization techniques and tools for large and complex health care data to facilitate timely decision-making and trend/pattern detection. A prototype system will be developed to test the effectiveness of this approach on a large-scale health care database that is currently available at Regenstrief Institute. More specifically, we want to develop a public health use case leveraging a Notifiable Condition Detector (NCD) dataset that contains reportable disease conditions that are transmitted to Indiana public health authorities (over 800,000 reports). Clinicians and public health stakeholders seek to uncover informative trends contained within the growing population-based datasets. To support knowledge discovery, we first extract meaningful terms and their associations and attributes from the raw data by applying data mining and text mining algorithms to construct a concept space. A browser-based user interface is developed to enable interactive online data exploration. A suite of visualization algorithms and techniques are developed and implemented within the prototype system.

KEYWORDS
Information visualization, Visual analytics, Public health data, Notifiable condition detector, Text mining, Data mining

OVERALL PROJECT SUMMARY

The project had four primary goals, all of which were accomplished.

I. Concept space definition
II. Algorithm design
III. System design and implementation
IV. System Prototyping and Usability Evaluation.

Concept Space Definition

The "concept space" represents a uniform layer of clinical observations and their associations and serves as a platform for users to explore the data using visualization and analysis methods. The concept terms are derived from data mining and text-mining processes applied to the use case datasets. For this project we focus on a population health use case that leverages an automated Notifiable Condition Detector (NCD). The NCD dataset contains 833,710 notifiable cases spanning more than 10 years from among 439,547 unique patients [1]. An additional dataset linked to the original NCD patient's data was extracted from the Indiana Network for Patient Care (INPC) health information exchange containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories [2]. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including Disease, Symptom, Mental behavior, and Risky Behavior. An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology.

Data Model Design

Considering the visualization-specific requirements, we designed a three-layer data model (Figure 1) to store the NCD and supporting text dataset. The first layer contains base tables for the entities included in our ontology: patient, disease, location, and other terms. The table for these additional terms has four subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains associations between the primary patient entity and additional three supporting entities. The third layer contains indirect associations between disease, term and location and was constructed using data mining techniques. Designs for the specific supporting schema and classes of associations were informed by the data mining results and the data elements necessary to support each specific visualization. Further, to avoid costly database scans during visualization execution the schema also includes pre-computed aggregate data necessary to support the specific visualizations. Pre-computed aggregate data include joint statistics such as entity association frequencies, e.g., the number of instances of "disease X" associated with "location Y".

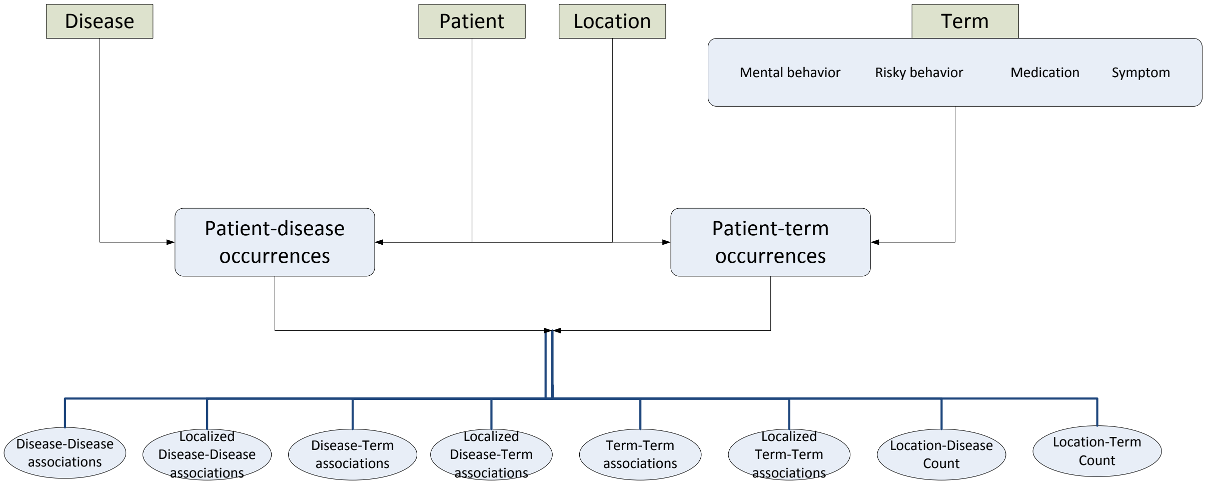

Disease
Patient
Location
Term
Mental behavior
Risky behavior
Medication
Symptom
Patient-disease occurrences
Patient-term occurrences
Disease-Disease associations
Localized Disease-Disease associations
Disease-Term associations
Localized Disease-Term associations
Term-Term associations
Localized Term-Term associations
Location-Disease Count
Location-Term Count

Figure.1 Database model

Data Cleansing

To preserve patient confidentiality we created a randomly assigned, pseudonymized patient identifier, (called “PseudoID”) linking records within and among the NCD and INPC datasets, but conveys no identifiable traits. In rare instances a PseudoID may falsely match more than one patient. To avoid this error, we used three additional fields to verify that records sharing the same PseudoID represent the same patient. The three additional fields are gender, race, and date of birth. If two records have the same PseudoID but disagree on non-null genders with values of ‘M’ or ‘F’, then the records are treated as separate patients. The race validation rule functions similarly to gender validation. For date of birth validation, we standardize date of birth format as “yyyy-mm-dd” and apply the longest common subsequence string comparator. If the ratio between the length of the longest common subsequence over the length of yyyy-mm-dd format is less than a certain threshold, date of birth validation fails. Records are determined to represent the same patient only if all three additional fields pass validation.

After cleansing, the database contained 439,547 patients, 1,976 diseases, 3,756 locations and 3,851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2847 medications). The second layer of the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. At least 90,376 patients are associating with at least one term non-disease. All of these patients have a least one disease. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term.

We removed duplicate public health case reports, which were defined as record having the same patient, the same date, and the same notifiable condition. We subsequently identified the most common reported conditions. The condition “Lead Exposure” was found among 256,823 patients. However, lead poisoning is not common in practice. “Lead Exposure” has the highest occurrence because Indiana’s reporting law requires that all laboratories performing blood lead tests report the results of those tests, whether normal or abnormal. Therefore, even when the test result is in the normal range, the test was reported. It leads to a high number of records on “Lead Exposure” in the data, while most of the report has negative results. Additional frequently reported conditions included: 1) Staphylococcus Methicillin-Resistant Aureus (MRSA), 2) HIV, 3) Chlamydia Infection, 4) Hepatitis B, 5) Hepatitis C, 6) Gonorrhea, 7) Chickenpox, 8) Measles, 9) Hepatitis A, 10) Enterococcus Vancomycin-Resistant (VRE) 11) Trichomoniasis 12) Syphilis. Figure 2 shows the number of occurrences of the most common diseases.

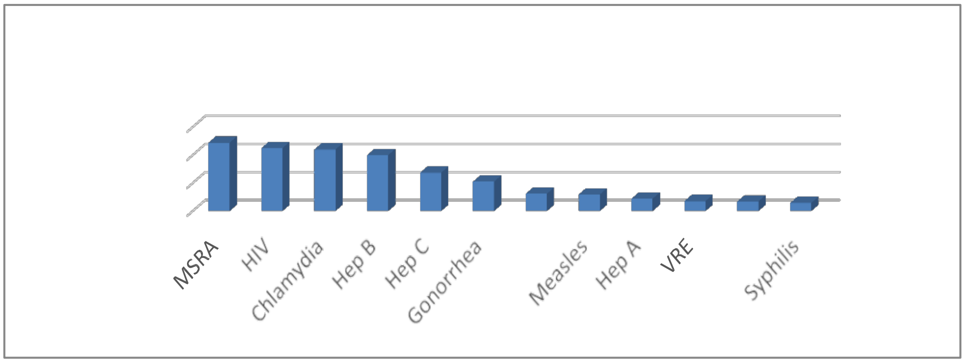


Figure 2: Most Common Diseases in the NCD dataset

We analyze the disease distribution across races. Here we compare the difference between the two largest races: white and black. The result is shown in Figure 3, with the black bar representing the occurrence percentage of each disease among black patients and the blue bar representing the occurrence percentage of disease among white patients. It shows that among black patients, CHLAMYDIA INFECTION and GONORRHEA are the most common conditions in the NCD data. TRICHOMONIASIS and SYPHILIS are also more common in black patients than in white patients. Among white patients, the most common condition is STAPHYLOCOCCUS METHICILLIN-RESISTANT aureus (MSRA).

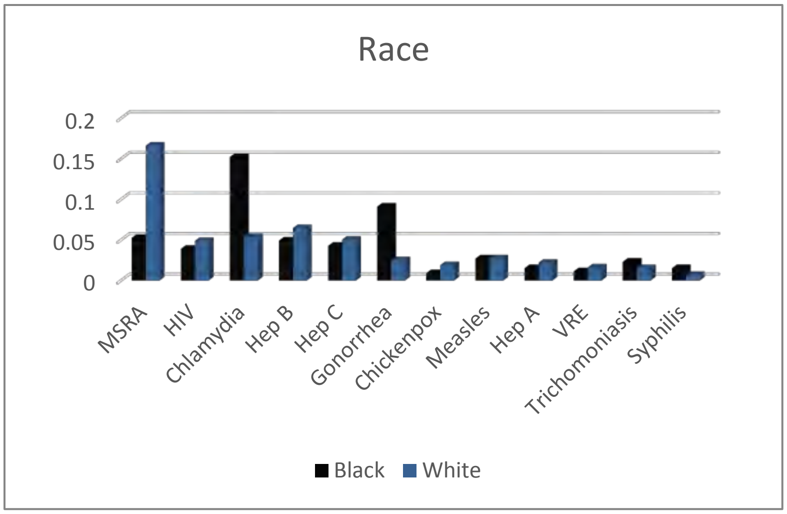


Figure 3: Diseases Distribution Across Race

Query efficiency

We test the database efficiency by three sets of common samples queries designed by visualization and health science experts. The first query set is about geographical distribution of one or a combination of diseases. The second query set retrieves strong associated diseases to a given disease. The third query set finds common diseases occurring at a given range of age. Table 1 summarizes the performance of three types of queries and suggests that the further optimization will be required for effective user interactions during interactive visualization.

Query set	Example	Involved tables	Runtime
1	Geographical (at city level) distribution of chlamydia	location, diseases, patient-disease occurrences	12s
2	List the diseases associating with chlamydia	diseases, associations	0.5s
3	What are the most common diseases for patient age from 20 to 40	diseases, patients, patient-disease occurrences	16s

Table 1: 3 query set for testing database

Algorithm Design

There are 3 types of algorithms that need to be developed: (1) Text Mining algorithms to extract concept terms from clinical notes; (2) Data mining algorithms, such as association mining and clustering; and (3) visualization algorithms for various visualization tools.

Text Mining

The NCD receives clinical data that includes the diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. But much of the information pertaining the patients' condition is available in the clinical reports. Mining these reports can provide a bigger picture of various other conditions that the patient experienced during his/her treatment. This information can provide valuable insights on the patients' socioeconomic condition, behavior risk factors, environmental factors and genetic information (family history). Natural Language Processing (NLP) provides a means to augment the NCD data analytics with the information discovered from these clinical reports.

Text Mining Process

NLP techniques were carried out to process 325791 clinical notes that contain patient discharge summaries, laboratory reports, patient history, etc. Although these records are de-identified due to which the patient specific information are absent, a pseudo-patient Id has been provided to help process the reports. Basic processing of the reports was performed for converting the clinical notes from XML format to simple text format and sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information from the reports. This was done with the help of UMLS [3] database which is a repository of clinical and health related terms. Once the entities were extracted using NER, negation analysis was applied using NEGEX algorithm [4] to remove negated terms. Figure 4 shows the process that was used in extracting this vital information from the reports.

The advantage of using UMLS is that all variations of clinical terms get captured that provide a large set of terms available for further analysis. For example, clinical notes that indicate “Hepatitis” contains terms like “Hepatitis”, “Hepatitis B”, “Hep”, “Hep B” etc. The large number of terms extracted contains different occurrences of the same diseases, symptoms, etc. We apply stemming and grouping algorithms to reduce the total number of terms. The identified terms are stored in different data tables and joined using the pseudo-patient Id.

Comorbidity Analysis

Once the data tables are constructed, we perform deeper analysis to compute the comorbid conditions of the diseases. For this, we use the *tf-idf* (term frequency – inverse document frequency) vector space model [5] to identify the significantly co-occurring diseases. The *tf-idf* model is considered to be an effective text mining model that provides the importance of a term/word to a document in a collection of documents. This model uses the concept of relevance and co-occurrence of terms. Equation (1) gives the relevance of a term *j* w.r.t. a document *i*,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j}) \quad (1)$$

where w_{ij} = relevance of term *j* in the patient record *i*; t_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; *N* = total number of records (*N*=325791).

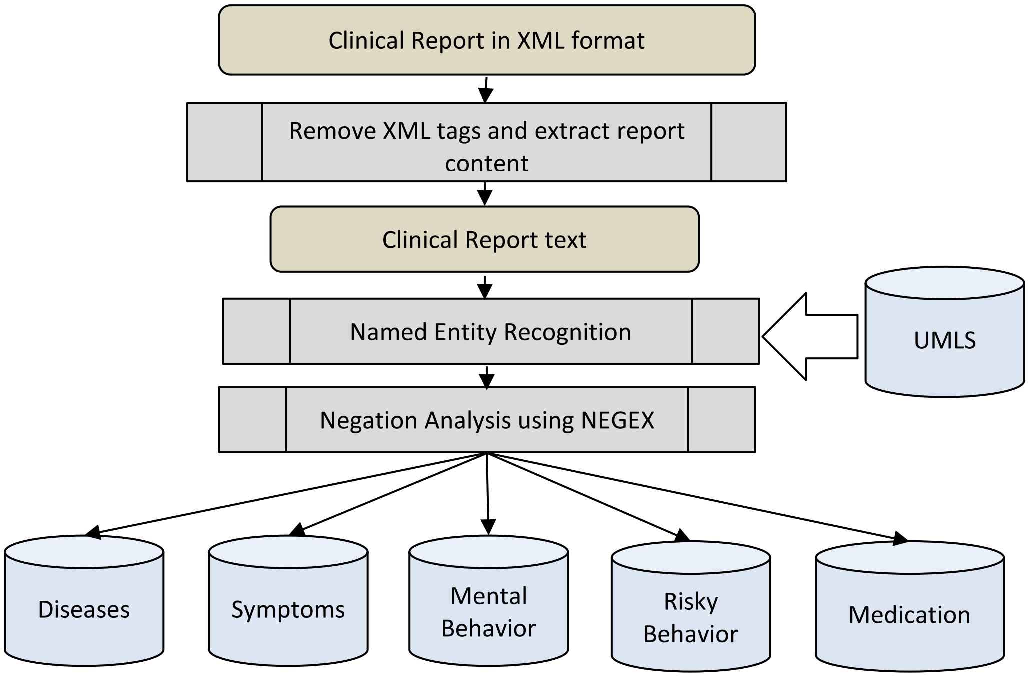


Figure 4. NLP steps applied on Clinical Reports

A particular term is more relevant *w.r.t*. a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight/score is attached with every association between a pair of terms [5]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k}) \quad (2)$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. The association score is 0 if the terms do not co-occur in any of the *N* records. Associations with non-zero scores are considered to be associated to the term.

After applying basic level processing on the reports, the clinical content from the reports was subjected to NER. UMLS was used for NER to identify the diseases, symptoms, mental behavior, risky behavior and medication terms from the 325791 reports. The total number of terms extracted for each category is given in Table 2. Figure 5 shows the most commonly occurring diseases with the number of reports in which they were found.

The top 10 diseases were analyzed using the *tf-idf* model to identify comorbidity of the diseases across the 325791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications) using Equation 2. Table 3 shows the top 10 diseases and the corresponding conditions.

Term Type	**Number of terms extracted using NLP**
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

Table 2: Total terms identified by NLP

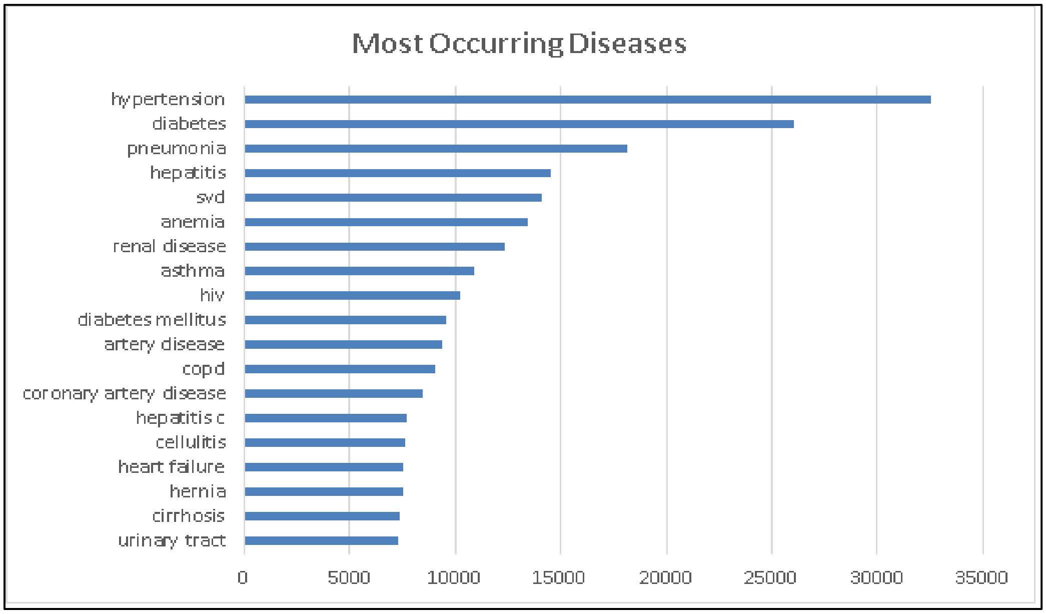
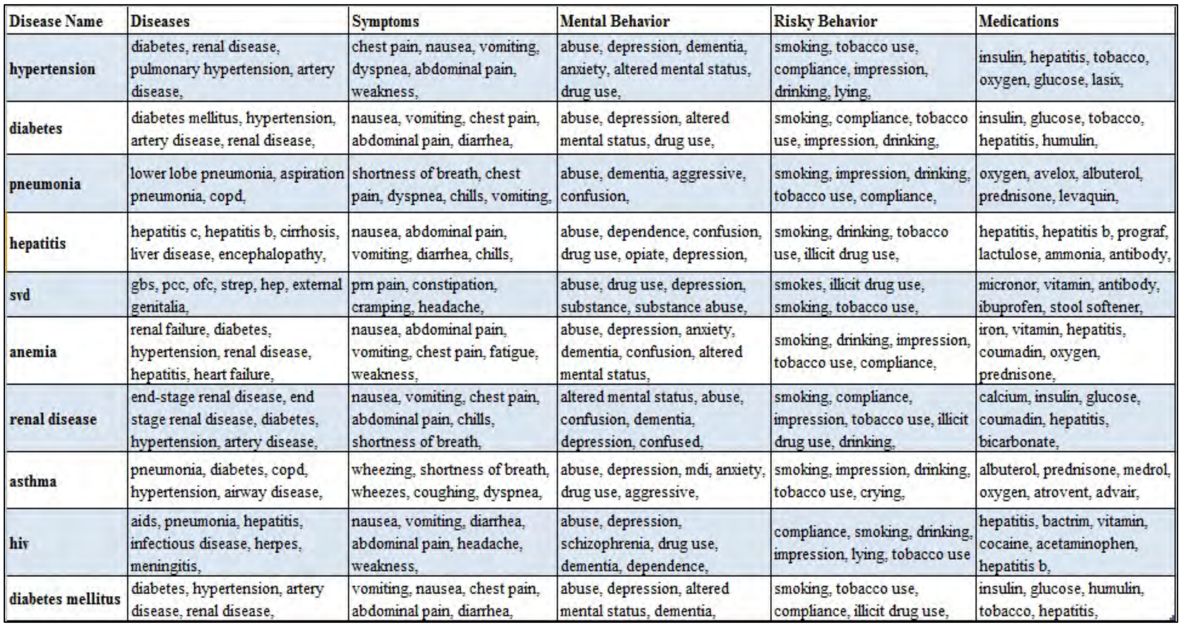
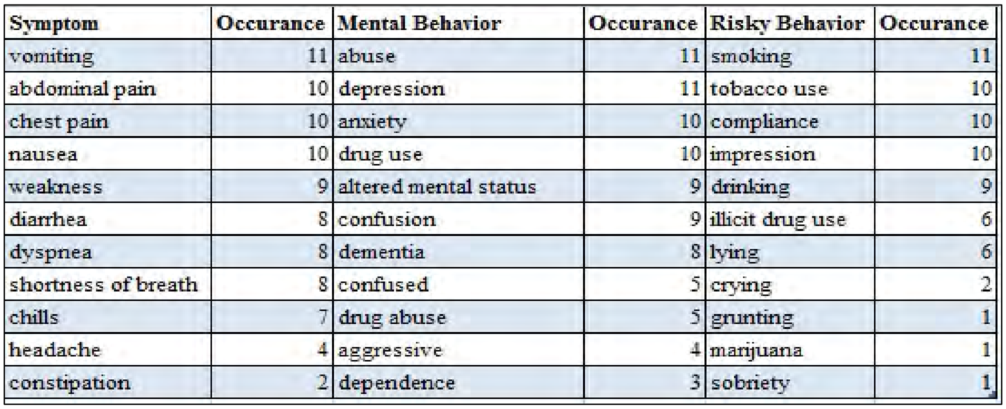


Figure 5: Most commonly occurring diseases and corresponding number of reports



Disease Name	Diseases	Symptoms	Mental Behavior	Risky Behavior	Medications
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion,	smoking, impression, drinking, tobacco use, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	prn pain, constipation, cramping, headache,	abuse, drug use, depression, substance, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal failure, diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease,	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused,	smoking, compliance, impression, tobacco use, illicit drug use, drinking,	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

Table 3**:** Comorbid conditions with top 10 most occurring diseases



Symptom	Occurance	Mental Behavior	Occurance	Risky Behavior	Occurance
vomiting	11	abuse	11	smoking	11
abdominal pain	10	depression	11	tobacco use	10
chest pain	10	anxiety	10	compliance	10
nausea	10	drug use	10	impression	10
weakness	9	altered mental status	9	drinking	9
diarrhea	8	confusion	9	illicit drug use	6
dyspnea	8	dementia	8	lying	6
shortness of breath	8	confused	5	crying	2
chills	7	drug abuse	5	grunting	1
headache	4	aggressive	4	marijuana	1
constipation	2	dependence	3	sobriety	1

Table 4**:** Most comorbid behaviors for the top 10 diseases

We also analyzed the most common conditions that occurred with these diseases (Table 4). It was interesting to find that well known behaviors such as “smoking”, “depression” and “tobacco use” were amongst the commonly occurring conditions.

Association mining

In layer 3, we compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities i and j, the rule confidence between *i* and *j* is computed as

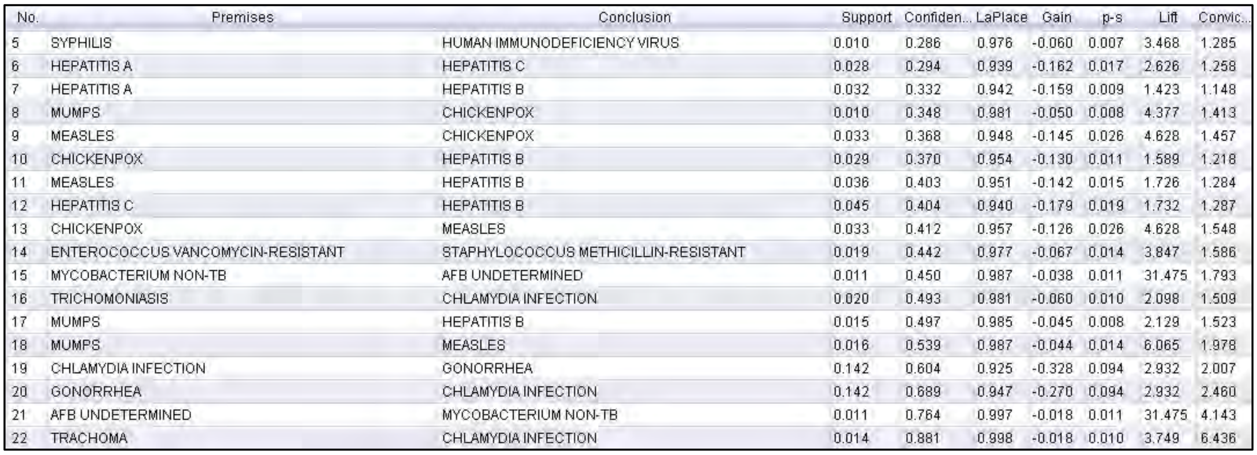
$$\text{Rule\_confidence}(i, j) = \frac{\left| i \wedge j \right|}{| i |}$$

in which $\left| i \wedge j \right|$ is the number of patients showing both entities *i* and *j* and | *i* | is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time

$$\text{Happen\_before}(i, j) = \frac{\left| i \text{ before } j \right|}{| i \wedge j |}$$

in which $\left| i \text{ before } j \right|$ is the number of showing *i* before showing *j*. We only compute localized association when then number of patients in the location is above 1000.

The database contains significant associations which are not widely reported in literature, such as Antidiarrheal treatment and runny nose symptom (confidence: 0.73), sclera and Tylenol treatment (confidence 0.70), posturing and Motrin treatment (confidence: 0.81), etc. Table 5 shows part of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the support, confidence, Laplace, Gain, p-s, lift and Conviction. We are working with domain experts on evaluating the association rules and tuning the parameters to produce optimum result.



No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convic...
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.626	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.368	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.954	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.987	-0.038	0.011	31.475	1.793
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.129	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.460
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Table 5: Association rules among diseases

Clustering analysis

We developed a co-clustering algorithm to cluster both diseases and text-mining terms to discover potential combinations of both diseases and terminologies, which could be disease subtypes or imply new biomedical patterns. The algorithm iteratively and partially [6] allocates the diseases or terms into clusters based on the rule confidence attributes. Let *K* be the number of clusters. To reallocate the diseases given the clustering allocation of terms, we select the *k* giving the maximum affinity score (*as* score) of disease *i* on cluster *k*. The *as* score is computed as

$$as\left(i,k\right)=\sum_{\forall j\in C_k}\left(p\left(i\mid j\right)-\underset{l}{\overline{p(l\mid j)}}\right)$$

in which C_k denotes the cluster *k*, *j* is the index of the term and $\underset{l}{\overline{p(l\mid j)}}$ is the mean of the associations given term *j*. $\underset{l}{\overline{p(l\mid j)}}$ is the repulse factor to prevent the case when all diseases and terms falls in one cluster. The process to reallocate the terms is similar to the diseases allocation process.

The algorithm can be executed in parallel by using a master-assistant computational model to improve efficiency. When reallocating diseases, the master routine sends the term-cluster allocations to all assistants and assigns the disease subsets for each assistant to reallocates. The assistants send the disease allocation results for the master routine for later use in terms allocation. We terminate the iterative allocation steps until the number of diseases/terms adopting new cluster is small and the clusters become stable.

We found several cluster containing close relationships between diseases and terminologies, such as {Biliary Sludge, HFA, Macrocytosis, Paroxysmal, Pseudogout, back discomfort, betimol, hesitancy, Intron A}, {Gastric Polyps, Kidney failure, antral, benefix, benicar}, {Duodenal Ulcer, Helicobacter Pylori, Malabsorpition, amylase, antimetics} and {appetite lost, immunoglobulin, retrovir} , etc. Some clusters highly correspond to specific diseases or medical processes. For example, appetite lost, immunoglobulin, retrovir are HIV related symptoms. Meanwhile, some clusters contain diseases and terms associated with several medical processes. For example, we found a cluster including gastrointestinal terms (Duodenal ulcer, Helicobacter pylori, Malabsorption, acyclovir, amylase and antiemetics), additive behavior (drinking, lortab andmarijuana) and cancer (methotrexate, vincristine and zofran). This cluster may suggest negative impact of addictive behavior toward digestive system. The appearance of cancer drugs in this cluster could raise a research question about the impact of additive behavior toward the metabolism process, which will further affect the cancer drug efficiency.

Sequential pattern mining

We construct the sequence of disease/term occurrence based on rule_confidence and happen_before association. Only associations with rule_confidence and happen_before association greater than certain threshold and covering at least 50 patients are included to construct the sequence and visualization. Due to the limited number of text records showing the test date, we only applied sequential pattern mining on disease association.

We found 105 disease-associations satisfying all 3 criteria about rule_confidence, begin_before_end and coverage to construct the frequent sequential disease patterns. We found 3 groups of sequences in the NCD data. The first group contains only one sequence Hyperplenism □ Annemia. Th
Staphylococcus Methicillin-resistant, Biliary Stricture, Cycsticercosis and Meconium Ileus, in which Fibrosis pulmonary and Meconium ileus stay at the triggering position. This disease group may raise additional research questions since these diseases occur at different organs. The last group is marked by Hepatitis A and 56 other diseases staying at the triggering position of Hepatitis A.

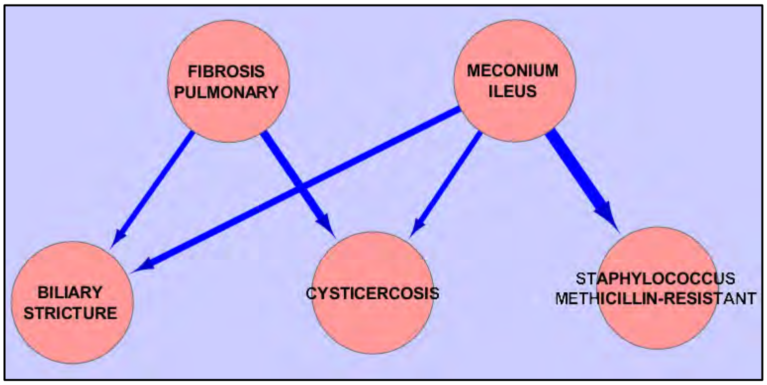


Figure 6. Fibrosis/Meconium-ileus sequence

Visualization algorithms

We have developed a suite of visualization algorithms for the interactive visual exploration of the health data represented in the concept space database.

Association graph

The opening visualization is an associative graph of the diseases and other terms from association mining. Association map is a graph visualization of the association relationships among the diseases and other terms in the concept space. It can serve as a platform supporting interactive selection of concepts to dynamically visualize data using a variety of tools in the visualization system. To draw an association graph, a spring-embedded algorithm [7] is used to layout the graph nodes by optimizing the following energy function:

$$E_s = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}k(d(i,j)-s(i,j))^2$$

Where *d(i,j)* is the 2D Euclidean distance of two nodes, and *s(i,j)* is a similarity metric of two nodes representing the heuristic of the layout. Edge thickness indicates the strength of association, and node size can reflects the number of other nodes to which a given node has a significant association, or the total occurrence of a term (e.g. disease) in the dataset. Nodes can be selected, and the graph will be quickly redrawn to only show other nodes which have significant association to the selected nodes.

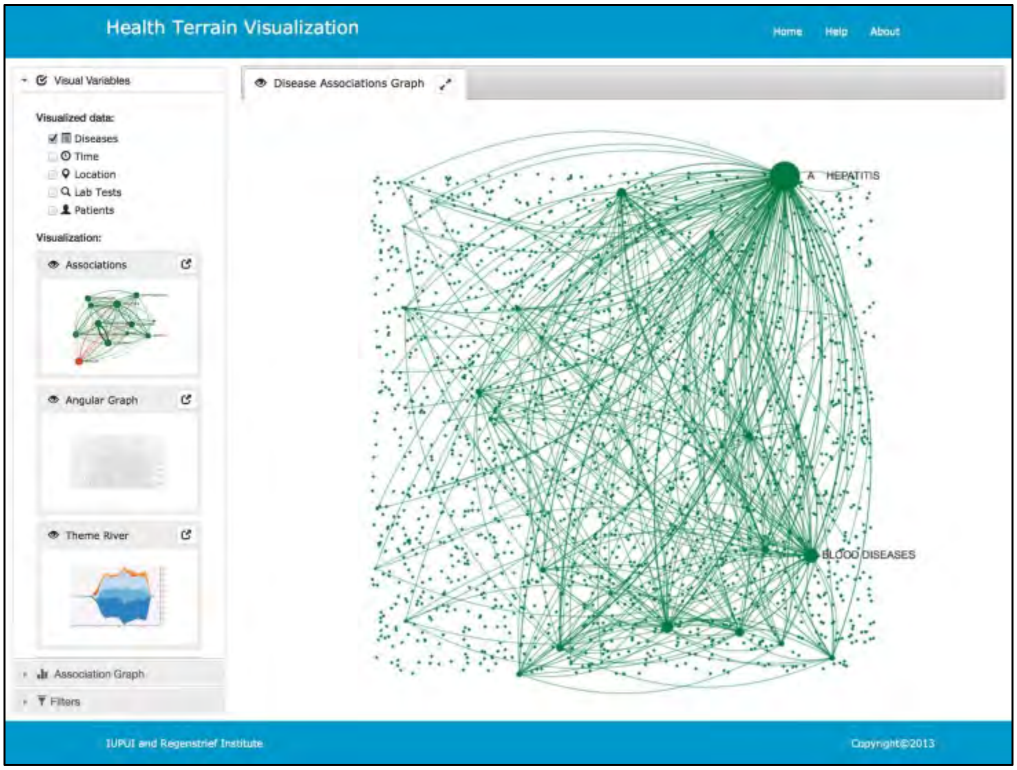
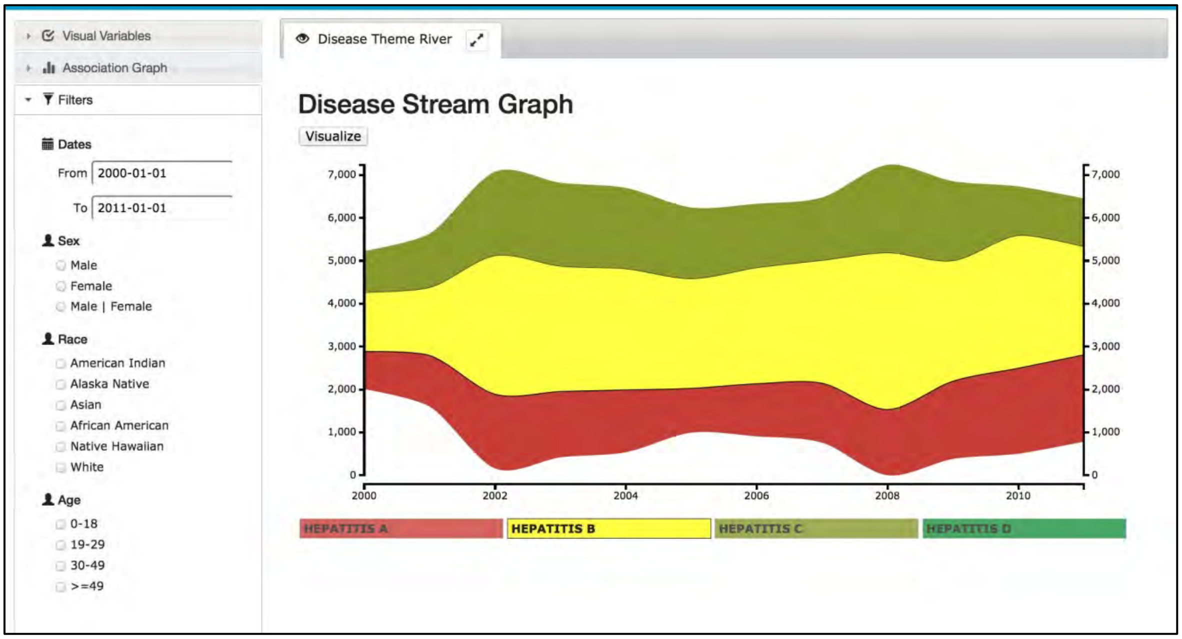


Figure 7. Disease association map and the web interface of the HealthTerrain system

Theme River

Theme river view [8] shows the aggregate trend for the terms (e.g. diseases and symptoms) selected by the user for a given time period. Each term (a theme) is visually represented as a river stream, and implemented as a filled curve plot along the horizontal time axis, with y-axis representing the occurrence of the term. Multiple themes are stacked together vertically for side-by-side comparison of the streams over time, as well as the possible interactions.


From 2000-01-01

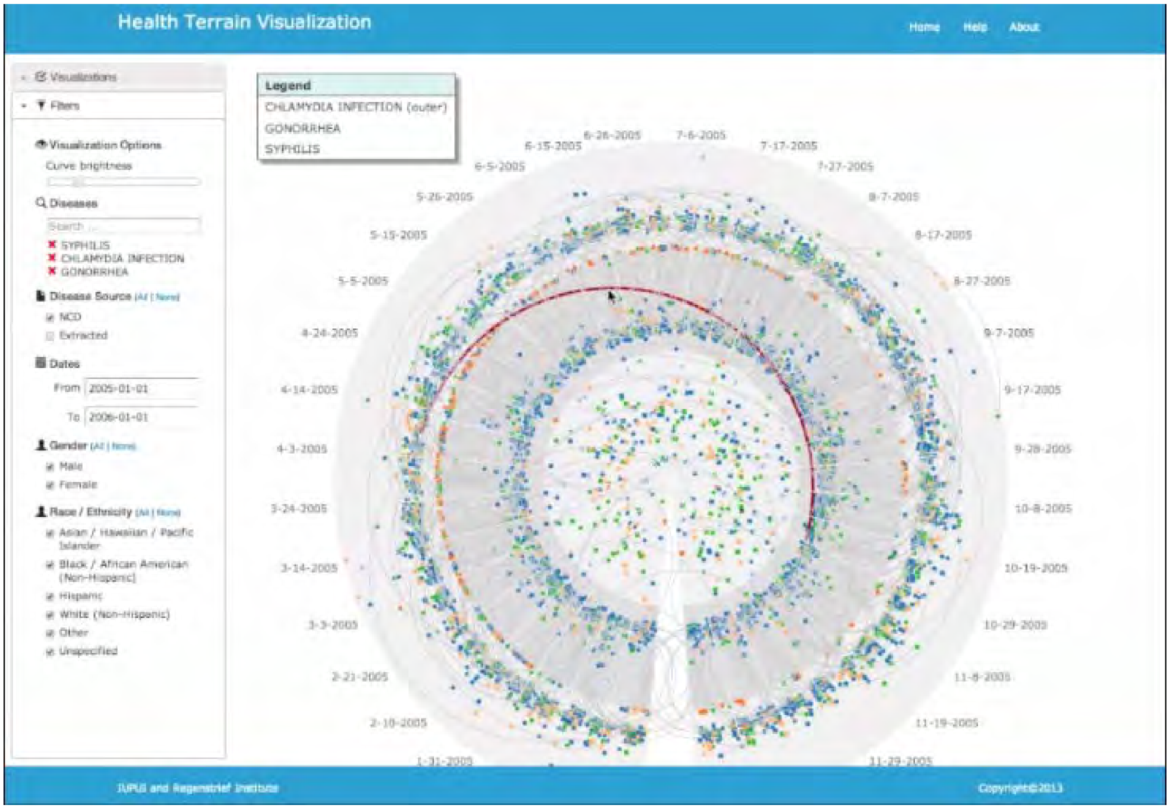
To 2011-01-01

Figure 8. The Theme River visualization for Hepatitis A, B, C and D

Ring graph

In order to view more detailed patient level data, we developed a new patient visualization method called Ring Graph. In Ring Graph, each patient is modeled as a point in a radial coordinate system. The radial space is subdivided into multiple rings, each of which represents one visualization term that was selected from the association map. These terms are typical disease names, but can also be other associated terms such as symptoms and risky behaviors. The circumference of this radial space represents the time-axis. Thus, time is encoded as the radial angle of the points (patients). Ring Graph shows the distribution of patient-level data over a time-attribute space. One significant attribute, for example "age", will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots.

Occurrences of the same patient associated with multiple terms (e.g. diagnosed with multiple diseases) are connected with curves across the graph. A connecting curve will be highlighted when there is mouse over on the patient or the curve. Details of a patient record can also be shown by mouse over. To avoid clutter, the connecting curves are drawn with adjustable semi-transparent lines. Lowering the transparency can reveal more clearly the associations between terms. Figure 9 show an example of the Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis over a time period.



Legend
CHLAMYDIA INFECTION (outer)
GONORRHEA
SYPHILIS

Diseases
✖ SYPHILIS
✖ CHLAMYDIA INFECTION
✖ GONORRHEA

Disease Source (All | None)
☑ NCD
☐ Extracted

Dates
From 2005-01-01
To 2006-01-01

Gender (All | None)
☑ Male
☑ Female

Race / Ethnicity (All | None)
☑ Asian / Hawaiian / Pacific Islander
☑ Black / African American (Non-Hispanic)
☑ Hispanic
☑ White (Non-Hispanic)
☑ Other
☑ Unspecified

Figure 9. A Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis

Texturization

In texturization, texture images are constructed to represent the overall data trends and distributions in different geospatial regions. Once the textures are generated, we will first visualize them on a 2D geographic map as a heatmap image, and then map them to terrain surfaces. There are two different types of textures that will be generated here (1) noise pattern texture for the representation of multiple attributes; and (2) offset contour texture for the time-varying data representation.

Noise Texture

We aim to represent multiple attributes for each geographic region using color coded texture patterns so that the users can easily perceive the representations of different attributes, not only within one region, but also its overall geospatial distributions across many regions in a geographic area (e.g. a state).

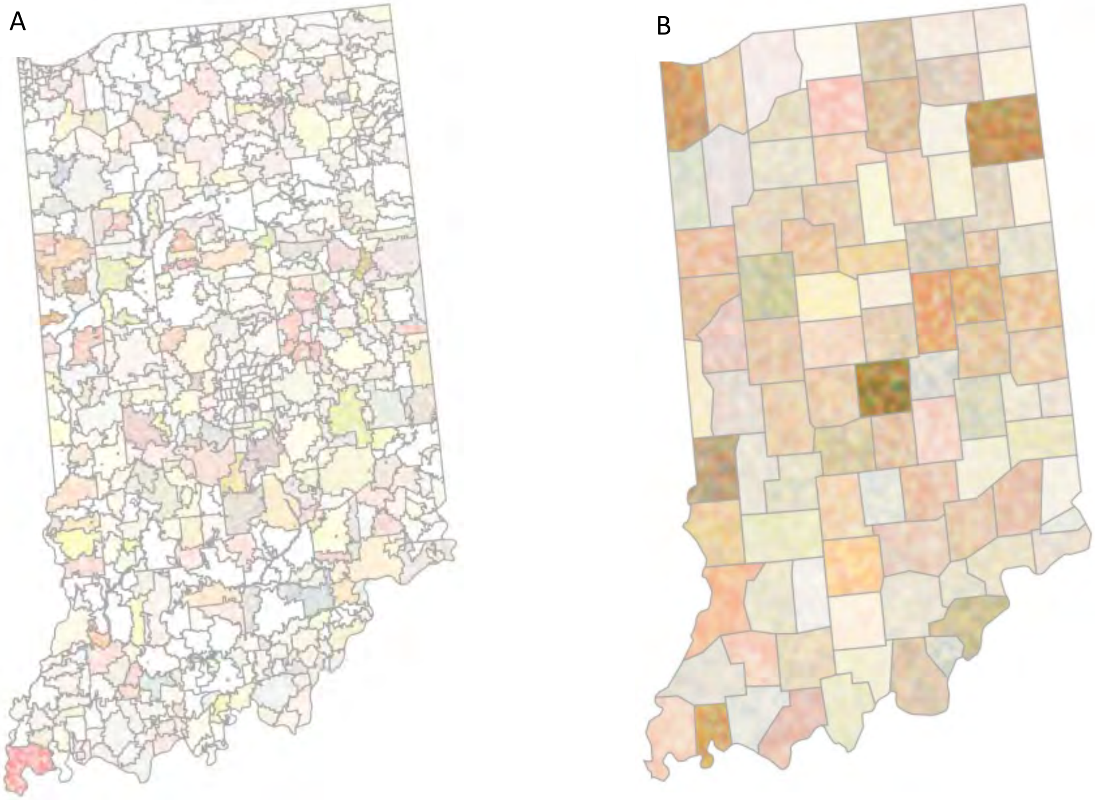


Figure 10. Heatmap views of noise textures over the Indiana state map: (a) county based; (b) zip-code based

We first construct noise patterns to create a random variation in color intensity, similar to the approach in [9]. Different color hues will be used to represent different types of attributes, for example the occurrences of different diseases. A turbulence function [10] will be used to generate the noise patterns of different frequencies (sizes of the sub-regions of the noise pattern). These multi-scale patterns may be applied to different scales of geographic areas (e.g. counties vs zip-codes). Since the noise pattern involves the mixing and blending of different color hues, we choose to use an RYB color model instead of RGB model, as proposed in [9], since RYB color model provides more intuitive representation of the weights of different colors after blending. Figure 10 shows two examples of the heatmap views of three diseases, Diabetes. Hepatitis B, and Chlamydia, over the Indiana state map.

<u>Offset Contouring</u>

Offset contouring is designed to represent attribute changes over time within a geographic region. It can also be used to represent multiple attributes. Similar to the Noise Pattern approach, we first construct a texture image using offset contour curves to form shape-preserving sub-regions, and then use varying color shades or hues to fill the sub-regions to represent the change of attribute values over time, or to simply fill the sub-regions with different color values to represent multiple attributes. The offset contours are generated by offsetting the boundary curve toward the interior of the region, creating multiple offset boundary curves (Figure 11).

Figure 11: Offset contouring, with multi-attribute coloring and time-series coloring

There are several offset curve algorithms available in curve/surface modeling. But since in our application, the offset curves do not need to be very accurate, we opt to use a simple image erosion algorithm [11] directly on the 2D image of the map to generate the offset contours. Figure 3b and 3c shows the color-filled sub-regions after offset contouring. In time-series data visualization, the time line can be divided into multiple time intervals and represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time. This approach, however, has two limitations. First, when the boundary shape of a region is highly concave, the image erosion technique sometimes does not generate clean offset contours. This usually can be corrected using a geometric offset curve algorithm such as the one in [12]. A second limitation of this approach is that it requires a certain amount of spatial area to layout the contours and color patterns. In public health data, however, these attributes are typically defined on geographic areas, which provide a perfect platform for texturization. Figure 12 shows a few examples of the heatmap views of offset contouring over the Indiana state map.

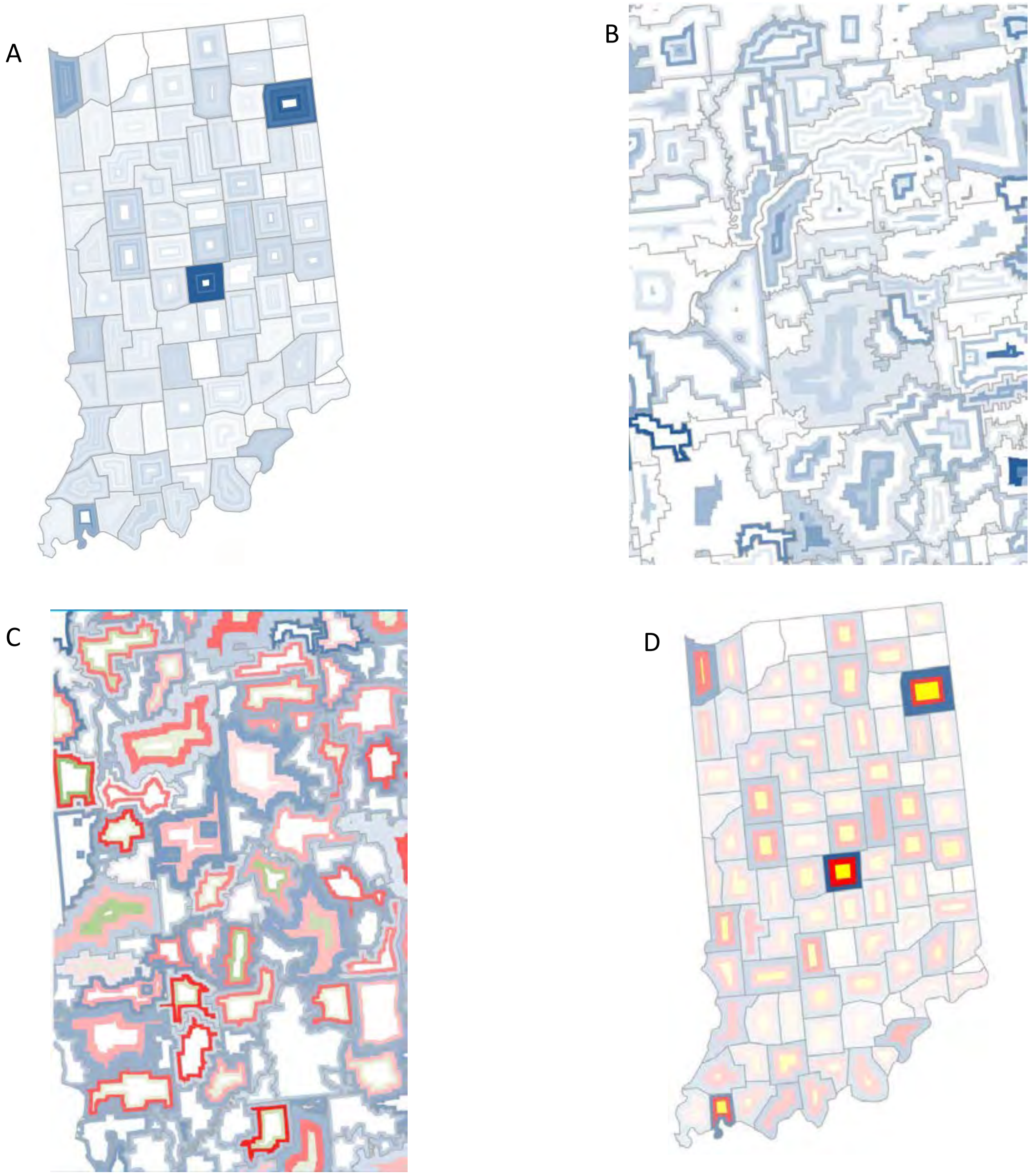


Figure 12. Heatmap views of offset contouring over the Indiana state map: (a) County based time-series data; (b) Zip-code based time-series data; (c) County based multi-diseases data; (d) Zip-code based multi-diseases data.

Texturized Terrain Surface

The heatmap views are effective in conveying the relative distributions of multiple attributes in different regions of a geographic area. The distribution of the total attribute values, however, becomes more difficult to perceive as the information has been disbursed by the texture patterns. This problem can be resolved by mapping the texture pattern onto a 3D terrain surface using the total attribute values as a height field.

A 3D surface can be constructed on top of a geographical region (e.g. the map of Indiana State). Typically, data are aggregated to individual geographical regions, such as counties and zip-codes, to form a height field. The height value can be, for example, the sum of multiple attribute values in a region, or the total occurrence of an attribute over the given time period for time-series data. To construct the surface, 3D scattered interpolation technique is applied so that every pixel point within the geographical boundary will have an interpolated height value. In our implementation, a Shepard interpolation method is applied:

$$d = \sum_{i=0}^{n-1} (1/r_i)^2 \cdot d_i \Big/ \sum_{i=0}^{n-1} (1/r_i)^2$$

where *d* is the height of an arbitrary point *P* within the geographical boundary, d_i are the known heights (attributes) at the known points C_i (*e.g*. center points of zip codes or counties), and r_i are the distances between *P* and C_i. A 2D image of the geographical map is used to limit the surface within the geographical border. This technique is implemented as a variation of our previous work on GeneTerrain [13].

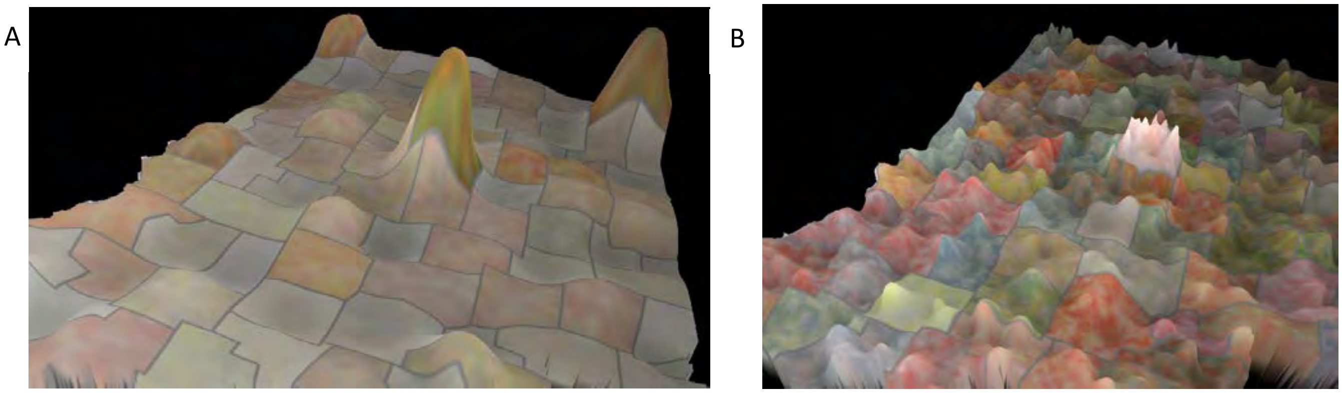


Figure 13. Terrain views of a multi-disease visualization over the Indiana state map. (a) County based textures and interpolation; (b) County textures and zip-code based interpolation.

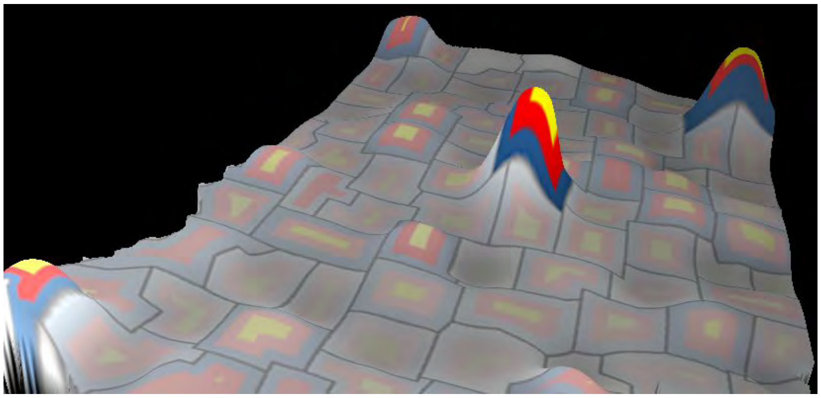


Figure 14. A terrain view of a county-based multi-diseases visualization

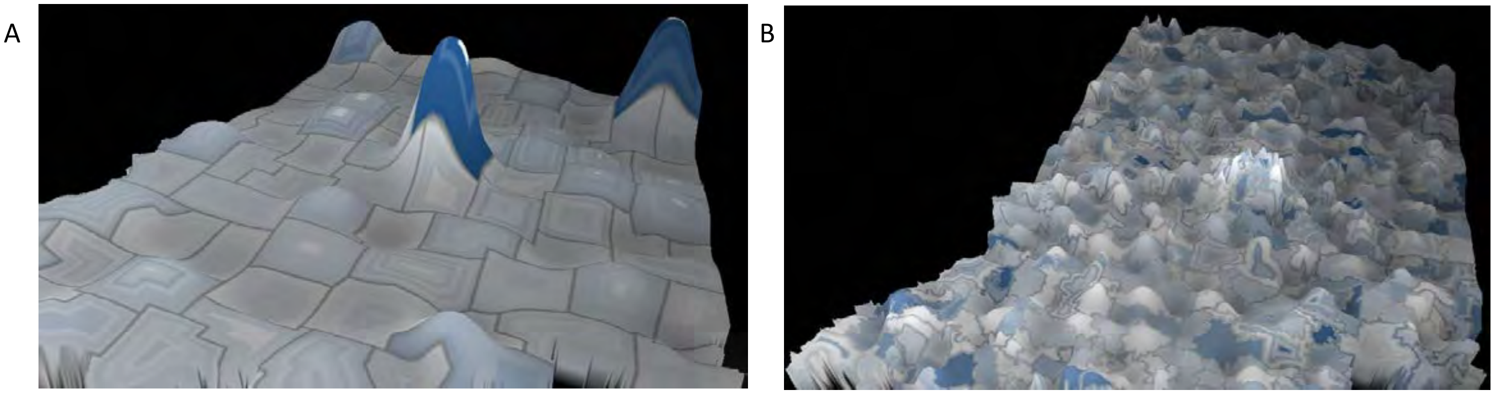


Figure 15. Terrain views of a time-series data over the Indiana state map. (a) County based; (b) zip-code based.

System Design and Implementation

We designed and implemented the initial structure and framework of HealthTerrain visualization. Initial plans had been to develop an installed executable application written in C++ and utilizing OpenGL for interactive visualizations. However, after some initial research and experimentation in the capabilities of modern web browsers (Google Chrome, Mozilla Firefox and Apple Safari) for 2D and 3D graphics, we came to the conclusion that WebGL in an HTML5 canvas would provide sufficient technical and graphical capabilities we need while appealing to a much broader potential user base with an established and maturing set of user experience patterns. Once focused on the web, we settled on an architecture pattern based primarily on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture. Ruby and Rails were picked as our server-side language and framework of choice for their elegant syntax, vibrant open source community, and ease of use.

The application itself is 3-part:

1. A MySQL relational database containing the results of offline text mining and statistical analysis research on the health data set provided to us by our partners at Regenstrief Institute.
2. A server-side RoR application for querying, modeling and manipulating data in the relational database.
3. An HTML/CSS/Javascript web GUI.

The user interface is a modern web GUI utilizing a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats such as Extensible Markup Language (XML) and JavaScript Object Notation (JSON). Interactivity is a primary goal as we seek to both visualize our data and provide opportunities for novel visual exploration and analysis.

The visualizations themselves utilize HTML, CSS, SVG, and WebGL technologies with a number of open-source Javascript libraries such as sigma.js, d3.js, jquery.js and three.js for drawing, displaying and interacting with the data and graphics. Figure 16 shows a screen shot of this interface that includes multiple visualization methods in a split window interface so that the visualization of the same dataset can be compared and analyzed.

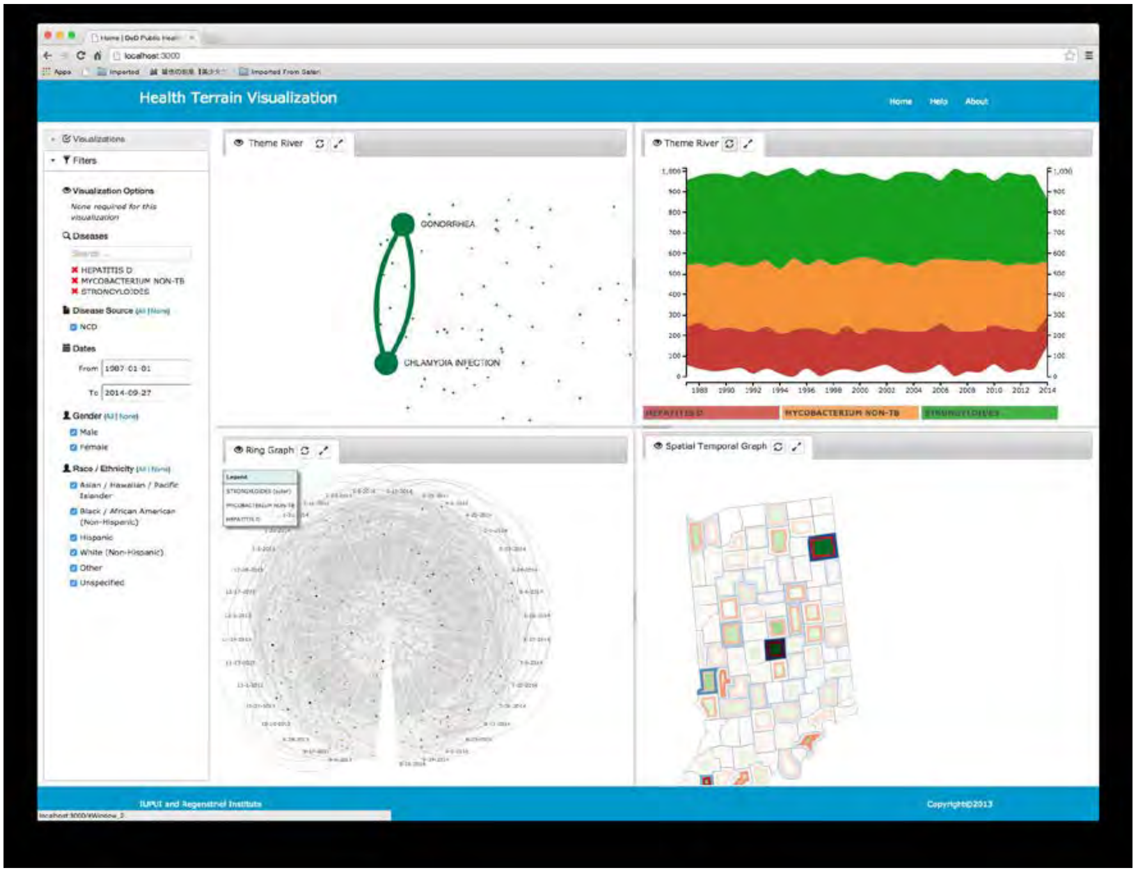


Figure 15. The web interface with split windows to show multiple visualizations of the same dataset.

System Prototyping and Usability Evaluation

After developing the data visualization framework, we imported de-identified communicable disease data and incorporated four operational visualizations including a network association graph, a ring graph, theme river graph, and 2D/3D cholorpleth (heatmap) spatiotemporal graphs. To perform a usability evaluation of this framework we recruited interviewees who represented potential end-users and visualization consumers, including public health epidemiologists with expertise in notifiable disease surveillance and syndromic surveillance; Indiana University faculty from the school of Public health; biomedical informaticians with public health informatics expertise from the Regenstrief Institute; clinical practitioners; and program managers with advanced training in public health management.

For our usability evaluation we adapted the National Institute of Standards and Technology (2007) definition of usability for our participants as the “effectiveness, efficiency, and satisfaction with which intended users can achieve their tasks and the intended context of product use.” Using an unstructured qualitative interview process, we explored dimensions of effectiveness (accuracy in completing tasks), efficiency (perceived time and effort in accomplishment of tasks), and satisfaction (subjective response to the application).

Prior to reviewing each of the four visualizations in successive order, the interviewees were oriented to the following dimensions of the application: 1) the overall screen layout and structure; 2) the ways in which users could navigate within a screen; 3) the ways in which users could navigate to other screens; 4) the ways in which users could navigate to the home screen; the ways in which users would move from field to field; and 5) a description of key commonly used buttons, icons, and links.

After presenting each visualization, the interviewees were asked to comment on the perceived dimensions of effectiveness, efficiency, and general satisfaction. Where necessary, exemplar leading questions were prepared to stimulate discussion, and included: “comment on your perceived satisfaction with the time required to interact with this visualization”; “how satisfied would you be with the perceived effort to interact with this visualization?”; “how confident are you that you could use this visualization to support your daily work flows on a routine basis?”; and “how quickly do you think most users would learn to perform the functions needed for this visualization?" The interviewees' responses were synthesized, stratified by each visualization and are summarized below:

Association Network Graph

As a general theme, the interviewees felt that including in-line guidance or pop-up descriptions (e.g., using mouse-overs) for each visualization parameter would provide end-users with valuable information to guide their use the tool. For example, the purpose of the association "threshold" parameter used in the association graph to create edges was unclear, and interviewees sought further definition. Interviewees noted that the visualization loading time, while less than 10 seconds, could be improved to enhance overall user satisfaction. The meaning of the colors of the edges in the graph was unclear, and interviewees felt they should be more clearly defined in the application. Public health stakeholders expressed the clear value of being able to quickly identify associations among multiple diseases, and they were pleased with the ability to filter out extraneous nodes and create sub-networks for strongly associated diseases. The interviewees felt that edges in the graph should contain metrics characterizing the strength of the association between nodes (disease).

Ring Graph

The interviewees described this visualization as being particularly complex and exhibiting high information density; some felt that the density obfuscated important information and were concerned that individual cases may be overlooked. The interviewees required substantial introduction to the graph prior to expressing recognition of the value of the visualization. Several commented that the extended (90-120 second) loading time was sub optimal, and hindered overall satisfaction, usability, and efficiency. While the dimensionalities of disease, age, gender, race, and time were generally perceived to be useful, the interviewees suggested that allowing those dimensions to be configurable would improve the utility of this visualization. One epidemiologist interviewee noted that their team likely would not use this visualization to identify disease outbreaks, but would instead use this visualization after an outbreak has been detected through other means in order to explore the relationships and characteristics of individuals within an outbreak in order to identify potential risk factors and target interventions. Another suggested that the circular format could be confusing and may obfuscate data; it was suggested that the graph be transformed into a linear format to potentially improve interpretability. One interviewee noted, "This graph has the potential to make me think about things that I wouldn't otherwise, and that has value to me."

Theme River Graph

Interviewees generally expressed that the theme river visualization provided a consumable, informative high-level comparison communicable disease incidence over time. Multiple interviewees indicated they would prefer case counts to begin at a common baseline on the y-axis; the variable heights and irregular sides of the theme river graph were felt to hinder interpretability. A consistent linear y-axis baseline of zero was felt to potentially enhance year-to-year comparisons over the default theme River visualization.

Spatiotemporal Graph

The three-dimensional version of this visualization was perceived to be more informative than the two-dimensional version. Commenters noted that the two-dimensional color variations within counties were challenging to interpret; the varying color intensity combined with varying band widths for each disease confused the interviewees. Some noted that continuous variation in color intensity may be less interpretable than dividing the range of disease incidence into a discrete set of ranges. Interviewees stated that presenting disease incidence as a three-dimensional height substantially improved interpretability and understanding of the data. There is wide variation in disease rates among counties (a small number of counties contain significant portions of overall disease); this variation obfuscates details in lower prevalence regions. Consequently the interviewees suggested that an additional feature enabling nonlinear scaling to highlight details in lower prevalence counties would be useful. They further suggested that presenting these data as incident rates (new cases per total population in the county) versus absolute counts (new cases) could improve interpretability and overall satisfaction. Epidemiologist interviewees requested extended functionality to visualize the highest prevalence diseases in each county.

General observations

Due to the data privacy policy provisions of the institutional review board research process, we used obfuscated de-identified clinical data for the usability assessment. The interviewees noted that further assessment of the usability of these different visualization tools would be enhanced by reviewing fully identified data rather than the de-identified obfuscated data currently used for research and development purposes.

KEY RESEARCH ACCOMPLISHMENTS

1) Designed and implemented a MySQL relational database, as a representation of the concept space, which is derived from the NCD dataset by data mining and text mining algorithms.
2) Natural Language Processing techniques were carried out to process 325791 clinical notes to extract new terms including diseases, symptoms, and mental and risky behaviors.
3) Data mining techniques were applied to extract associations between terms in the concept space, and to discover new cluster terms.
4) Designed and implemented a suite of interactive visualization algorithms that allows the users to interactively explore the data based on the user selected terms and filters. These include the association map, the theme river graph, the ring graphs, and the texturization based spatiotemporal visualization. The ring graph and the texturization based spatiotemporal visualization are two novel techniques that has never been developed by others.
5) Designed and implemented a web based graphical user interface for the prototype system, and successfully integrated the programming interfaces between the user interface, visualization, and the database.
6) Designed and tested an evaluation procedure for health data visualization system.

CONCLUSION

We have made significant accomplishments in this project, including: (1) created a concept space definition, which represents a schema tailored to support diverse visualizations and provides a uniform ontology that allows the system to be leveraged for many types of health care datasets through individually designed text and data mining procedures; (2) designed and implemented a suite of novel visualization algorithm, as well as data and text mining analytics techniques; (3) developed a prototype visualization system for the exploration of large-scale, real-world health data; and (4) Designed and tested an evaluation procedure for health data visualization systems. These components are integrated in a generalizable browser-based graphical interface, which enables flexible and free-form data exploration and hypothesis discovery. The system has received favorable initial feedback from users, and we believe it has potential as an open source tool to support health data visualization tasks.

PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS

Publications/Abstracts

Shiaofen Fang, Mathew Palakal, Yuni Xia, Sam Bloomquist, Thanh Minh Nguyen, Anand Krishnan, Shenghui Jiang, Weizhi Li, Jeremy Keiper, and haun Grannis. Health-Terrain: A Visual Analytics System for Health Data. Journal of American Medical Informatics Association, Accepted, 2014.

Shiaofen Fang, Shenhui Jiang, Sam Bloomquist, Mathew Palakal, Yuni Xia, Li Huang, Jeremy Keiper, and Shaun Grannis. Visualizing Large Healthcare Data by Geospatial Texturization. Submitted.

Presentations

J. Keiper, Shiaofen Fang, Yuni. Xia, Mathew Palakal, R. Shaun Grannis, Thanh Minh Nguyen, Sam Bloomquist, Anand Krishnan, Weizhi Li. A Public Health Data Visualization System Demonstration, AMIA 2014, Demo paper, Washington DC. Nov. 2014.

J. Keiper, Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Use Cases for Public Health Data Visualization. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Data Exploration of a Notifiable Condition Detector System. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC, Oct. 2013.

M. Palakal, S. Fang, Y. Xia, S. Grannis, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper. Detecting Comorbidity of Chlamydia from Clinical Reports. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

INVENTIONS, PATENTS AND LICENSES

Nothing to report.

REPORTABLE OUTCOMES

1. We have constructed 5 ontologies, one for each category - Disease, Symptom, Mental behavior, Risky Behavior and Medication based on the data extracted from the clinical notes. This helps us in eliminating noise and providing structure to our findings.
2. We have built a specialized database for the storage and real time query of the concept space and the NCD dataset. The database is general enough to be adapted to any other health care data and extensions of the concept space.
3. We have developed a prototype visualization system for healthcare data. The system provides a web based user interface that allows interactive visualization and exploration of a given dataset representing a use case.

OTHER ACHIEVEMENTS

Nothing to report

REFERENCES

Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/

Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html

B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11

Chapman , W.; Bridewell , ; Hanbury , ; Cooper , G. F.; Buchanan , G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* **2001,** *34* (5), 301–310.

Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342

Van Mechelen I, Bock HH, De Boeck P. Two-mode clustering methods:a structured overview. Statistical Methods in Medical Research 13 (5): 363–94, 2004

Stephen G. Kobourov. Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv: 1201.3011.

S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *Visualization and Computer Graphics, IEEE Transactions*, vol. 8, pp. 9-20, 2002

Nathan Gossett, Baoquan Chen. Paint Inspired Color Mixing and Compositing for Visualization. IEEE Symposium on Information Visualization 2004. 113-117

Ken Perlin. An image synthesizer. In Proceedings of SIGGRAPH85, pages 287–296. ACM Press, 1985.

Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.

Hoschek, J., (1988), “Spline Approximation of Offset Curves,” Computer Aided Geometric Design, Vol. 5, pp. 33–40.

You, Q., Fang, S., Chen, J. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks. Journal of Information Visualization, 2010; 9:1, 1-12.

APPENDICES

Attach all appendices that contain information that supplements, clarifies or supports the text. Examples include original copies of journal articles, reprints of manuscripts and abstracts, a curriculum vitae, patent applications, study questionnaires, and surveys, etc.

A Knowledge Discovery System for Notifiable Condition Detector Data

Thanh Minh Nguyen[1] , Anand Krishnan[2], Sam Bloomquist[1], Jeremy Keiper[3], Weizhi Li[2], Shaun Grannis, MD[3], Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2],
[1]Department of Computer Science, Indiana University – Purdue University, Indianapolis; [2]School of Informatics and Computing, Indiana University – Purdue University, Indianapolis; [3]Regenstrief Institute, Indianapolis, IN

Abstract

The Notifiable Condition Detector (NCD) system is an automated electronic lab reporting (ELR) and case-notification system developed by Regenstrief Institute. It has been used in Indiana for over ten years to report laboratory results for the detection of notifiable conditions such as novel H1N1 influenza, sexually transmitted diseases, lead poisoning, and salmonella[1]. In this paper, we present a knowledge discovery system which integrates database, text mining and data mining on the NCD data. We show that the knowledge discovery system could not only discover new associations among terminologies in health science but also enhance friendly visualization application to show more significant and useful data to users.

Introduction

The Regenstrief Institute implemented and maintains an HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in Indiana State Marion County. The Notifiable Condition Detector (NCD) System uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC). The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations[2].

The NCD, now operational in Indiana, automatically detects positive cases of indicated conditions and forwards alerts to local and state health departments for review and possible follow up. These alerts assist public health agencies to perform population health monitoring more efficiently and effectively.

Data Summary

The dataset we received from Regenstrief NCD system contains 833,710 observations. The dataset has been de-identified, with patient age and zip code pseudonymized. The dataset contains 47 columns including patient pseudo ID, condition name, test result name, test result value, test normal range, patient race, patient gender, etc. The missing data rate for columns varied substantially from 0% for column Patient Pseudo_ID to over 70% for column Test_Abnormal_Flag.

Methods

The NCD receives clinical data that includes the diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. But much of the information pertaining the patients' condition is available in the clinical reports. Mining these reports can provide a bigger picture of various other conditions that the patient experienced during his/her treatment. This information can provide valuable insights on the patients' socioeconomic condition, behavior risk factors, environmental factors and genetic information (family history). Natural Language Processing (NLP) provides a means to augment the NCD data analytics with the information discovered from these clinical reports.

Text Mining Process

NLP techniques were carried out to process 325791 clinical notes that contain patient discharge summaries, laboratory reports, patient history, etc. Although these records are de-identified due to which the patient specific information are absent, a pseudo-patient Id has been provided to help process the reports. Basic processing of the reports was performed for converting the clinical notes from XML format to simple text format and sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information from the reports. This was done with the help of UMLS[3] database which is a repository of clinical and health related terms. Once the entities were extracted using NER, negation analysis was applied using NEGEX algorithm[4] to remove negated terms. Figure 1 shows the process that was used in extracting this vital information from the reports.

The advantage of using UMLS is that all variations of clinical terms get captured that provide a large set of terms available for further analysis. For example, clinical notes that indicate “Hepatitis” contains terms like “Hepatitis”, “Hepatitis B”, “Hep”, “Hep B” etc. The large number of terms extracted contains different occurrences of the same diseases, symptoms, etc. We apply stemming and grouping algorithms to group these terms to reduce the total number of terms. The identified terms are stored in different data tables and joined using the pseudo-patient Id.

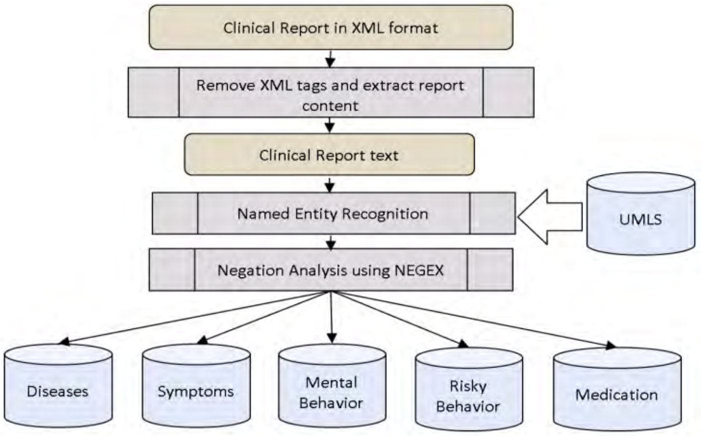


Figure 1. NLP steps applied on clinical reports

Comorbidity Analysis

Once the data tables are constructed, we perform deeper analysis to compute the comorbid conditions of the diseases. For this, we use the *tf-idf* (term frequency – inverse document frequency) vector space model[5] to identify the significantly co-occurring diseases. The *tf-idf* model is considered to be an effective text mining model that provides the importance of a term/word to a document in a collection of documents. This model uses the concept of relevance and co-occurrence of terms. Equation (1) gives the relevance of a term *j* w.r.t. a document *i*,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j}) \tag{1}$$

where w_{ij} = relevance of term *j* in the patient record *i*; t_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; *N* = total number of records (*N*=325791).

A particular term is more relevant *w.r.t.* a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight/score is attached with every association between a pair of terms[6]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k}) \tag{2}$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. The association score is 0 if the terms do not co-occur in any of the *N* records. Associations with non-zero scores are considered to be associated to the term.

Database design

Considering the visualization specific requirements, we designed a 3-layer database model (Figure 2) to store the NCD and supported text dataset. The first layer contains 4 base tables for 4 entities: patient, disease, term and location. The term table has 4 subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains the associations between the patient entity and other entities. Therefore, the NCD and supported text dataset could be recovered by joining tables in the second layer over the patient table. The third layer contains indirect associations between disease, term and location. Constructing the third layer require data mining techniques. We decide the specific tables and types of associations based on frequent queries requested from the visualization application. In addition, there are tables to store query results such as counting the number of patients having disease x at location y to avoid database scan during visualization execution. By caching the results of queries, the database can avoid having to repeat the potentially time-consuming and intensive operations (for example, sorting/aggregation, joins etc) that generated the query results and speed up rendering and visualization. The association could be both at global scale and at local scale.

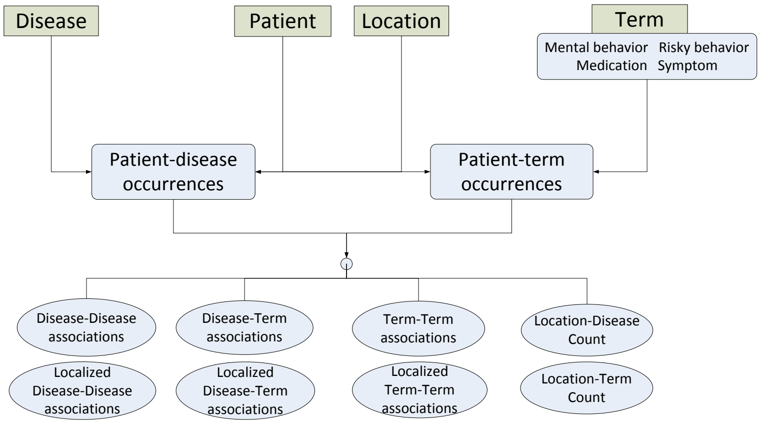

Disease
Patient
Location
Term
Mental behavior Risky behavior
Medication Symptom
Patient-disease occurrences
Patient-term occurrences
Disease-Disease associations
Disease-Term associations
Term-Term associations
Location-Disease Count
Localized Disease-Disease associations
Localized Disease-Term associations
Localized Term-Term associations
Location-Term Count

Figure 2. Database model

Data cleansing

There are cases in which one PseudoID may match to more than one patient. To avoid this error, we create three checkpoints to ensure that patients sharing the same PseudoID are indeed the same patient. Three checkpoints are gender, race, and date of birth. For gender checkpoints, if two patients show different genders in the records and none of the genders are NULL or U (unknown), then the gender checkpoint fails. The race checkpoint works in a similarly way. For the date of birth checkpoint, we convert the date of birth into yyyy-mm-dd format and perform longest common subsequence matching[7]. If the ratio between the length of the longest common subsequence over the length of yyyy-mm-dd format is less than a certain threshold, we will decide that the date of birth checkpoint fail. Different entries are from the same patient if and only if all three checkpoints pass.

In addition, we discovered cases when the terms mined from the text having different canonical name but representing the same entity, such as ‘pain’, ‘pains’ and ‘x_pains’. We applied the procedure similar to patient data cleansing to eliminate duplications among these terms.

Association Mining

In layer 3, we compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities i and j, the rule confidence between *i* and *j* is computed as

$$\text{Rule\_confidence}(i, j) = \frac{\left|i \cap j\right|}{|\,i\,|} \; 9$$

in which $\left|i \cap j\right|$ is the number of patients showing both entities *i* and *j* and $|\,i\,|$ is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time

$$\text{Happen\_before}(i, j) = \frac{\left|i \text{ before } j\right|}{|\,i \cap j\,|}$$

in which $\left|i \text{ before } j\right|$ is the number of showing *i* before showing *j*. We only compute localized association when then number of patients in the location is above a minimal threshold. The reason of enforcing the minimal patient threshold is to ensure the statistical significance of the associations identified.

Clustering Analysis

We develop a co-clustering algorithm to cluster both diseases and text-mining terms to discover potential combinations of both diseases and terminologies, which could be disease subtypes or imply new biomedical patterns. The algorithm iteratively and partially[10] allocates the diseases or terms into clusters based on the rule confidence attributes. Let K be the number of clusters. To reallocate the diseases given the clustering allocation of terms, we select the k giving the maximum affinity score (*as* score) of disease i on cluster k. The *as* score is computed as

$$as(i,k) = \sum_{j \in C_k} \left(p(i \mid j) - \overline{p(l \mid j)}_{l} \right)$$

in which C_k denotes the cluster k, j is the index of the term and $\overline{p(l \mid j)}_{l}$ is the mean of the associations given term j. $\overline{p(l \mid j)}_{l}$ is the repulse factor to prevent the case when all diseases and terms falls in one cluster. The process to reallocate the terms is similar to the diseases allocation process.

The algorithm can be executed in parallel by using a master-assistant computational model and. When reallocating diseases, the master routine sends the term-cluster allocations to all assistants and assigns the disease subsets for each assistant to reallocates. The assistants send the disease allocation results for the master routine for later use in terms allocation. We terminate the iterative allocation steps until the number of diseases/terms adopting new cluster is small and the clusters become stable.

Sequential Pattern Mining

We construct the sequence of disease/term occurrence based on rule_confidence and happen_before association. Only associations with rule_confidence and happen_before association greater than certain threshold and covering at least 50 patients are included to construct the sequence and visualization. Due to the limited number of text records showing the test date, we only applied sequential pattern mining on disease association.

Results

Text Mining Results

After applying basic level processing on the reports, the clinical content from the reports was subjected to NER. UMLS was used for NER to identify the diseases, symptoms, mental behavior, risky behavior and medication terms from the 325791 reports. The total number of terms extracted for each category is given in Table 1. Figure 3 shows the most commonly occurring diseases with the number of reports in which they were found.

The top 10 diseases were analyzed using the tf-idf model to identify comorbidity of the diseases across the 325791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications) using Equation 2. Table 2 shows the top 10 diseases and the corresponding conditions.

Table 1: Total terms identified by NLP

Term Type	**Number of terms extracted using NLP**
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

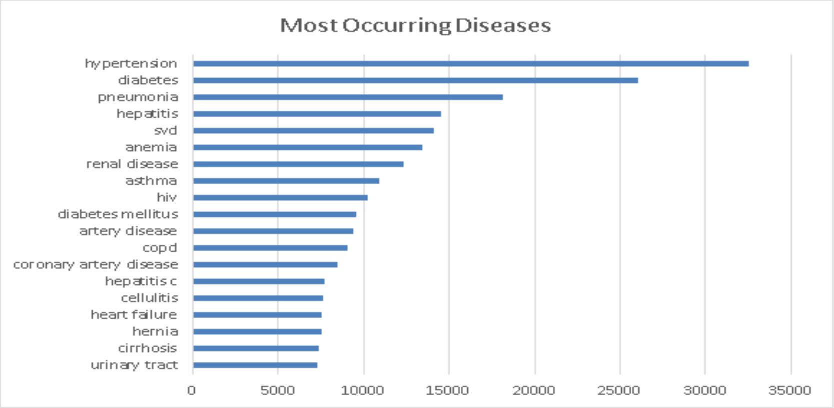
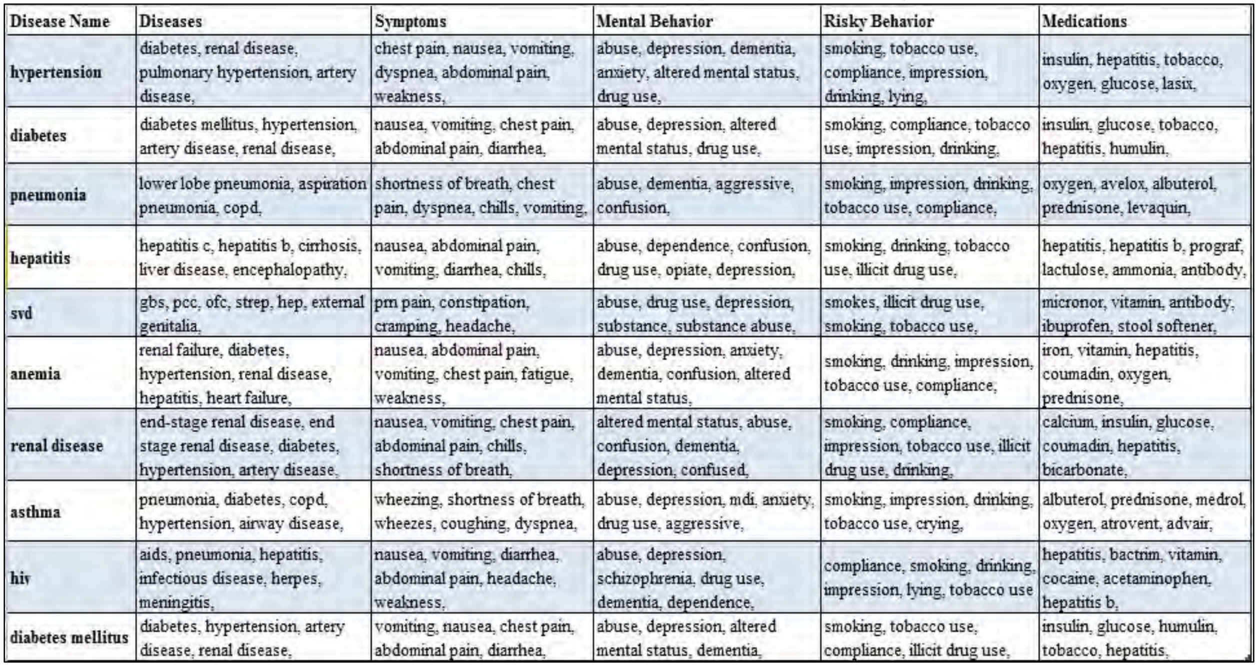


Figure 3. Most commonly occurring diseases and corresponding number of reports

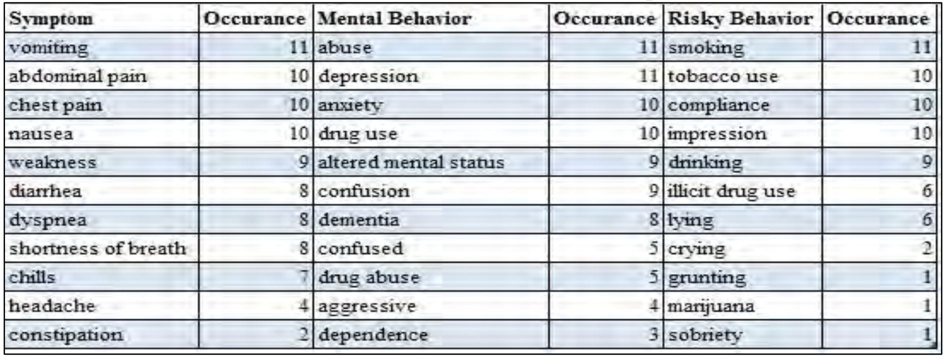
We also analyzed the most common conditions that occurred with these diseases. It was interesting to find (Table 3) that well known behaviors such as "smoking", "depression" and "tobacco use" were amongst the commonly occurring conditions.

Table 2: Comorbid conditions with top 10 most occurring diseases



Disease Name	Diseases	Symptoms	Mental Behavior	Risky Behavior	Medications
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion,	smoking, impression, drinking, tobacco use, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	pm pain, constipation, cramping, headache,	abuse, drug use, depression, substance, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal failure, diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease,	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused,	smoking, compliance, impression, tobacco use, illicit drug use, drinking,	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

Table 3: Most comorbid behaviors for the top 10 diseases



Symptom	Occurance	Mental Behavior	Occurance	Risky Behavior	Occurance
vomiting	11	abuse	11	smoking	11
abdominal pain	10	depression	11	tobacco use	10
chest pain	10	anxiety	10	compliance	10
nausea	10	drug use	10	impression	10
weakness	9	altered mental status	9	drinking	9
diarrhea	8	confusion	9	illicit drug use	6
dyspnea	8	dementia	8	lying	6
shortness of breath	8	confused	5	crying	2
chills	7	drug abuse	5	grunting	1
headache	4	aggressive	4	marijuana	1
constipation	2	dependence	3	sobriety	1

Database after cleansing

After cleansing, the database has 439,547 patients, 1976 diseases, 3756 locations and 3851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2847 medications). At the second layer, the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. However, there are only 90,376 patients associating with at least one term. All of these patients have a least one disease. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term.

We first remove from the dataset duplicate rows, where the same patient reports the same condition multiple times. Then, we identified the most common diseases in the data. "Lead Exposure" is the condition that has the highest occurrence- it occurs in 256823 patients; however, lead poisoning is not common in practice. The reason "Lead Exposure" has the highest occurrence in the data is that under the state's reporting law, all laboratories performing blood lead tests are required to report the results of those tests. Therefore, even when the test result is in the normal range, the test was reported. It leads to a high number of records on "Lead Exposure" in the data, while most of the report has negative results. Other than "Lead Exposure", the most common diseases include 1)Staphylococcus Methicillin-Resistant Aureus (MRSA), 2)HIV, 3)Chlamydia Infection, 4)Hepatitis B, 5)Hepatitis C, 6)Gonorrhea, 7)Chickenpox 8)Measles 9)Hepatitis A 10) Enterococcus Vancomycin-Resistant (VRE) 11) Trichomoniasis 12) Syphilis. Figure 4 shows the number of occurrences of the most common diseases.

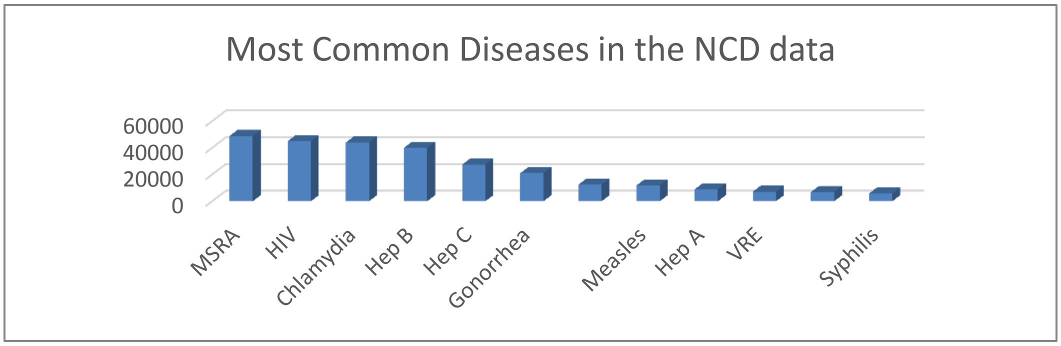


Figure 4. Most Common Diseases in the NCD dataset

Disease Distribution Across Race

We analyze the disease distribution across races. Here we compare the difference between the two largest races: white and black. The result is shown in Figure 5, with the black bar representing the occurrence percentage of each disease among black patients and the blue bar representing the occurrence percentage of disease among white patients. It shows that among black patients, CHLAMYDIA INFECTION and GONORRHEA are the most common conditions in the NCD data. TRICHOMONIASIS and SYPHILIS are also more common in black patients than in white patients. Among white patients, the most common condition is STAPHYLOCOCCUS METHICILLIN-RESISTANT AUREUS (MSRA).

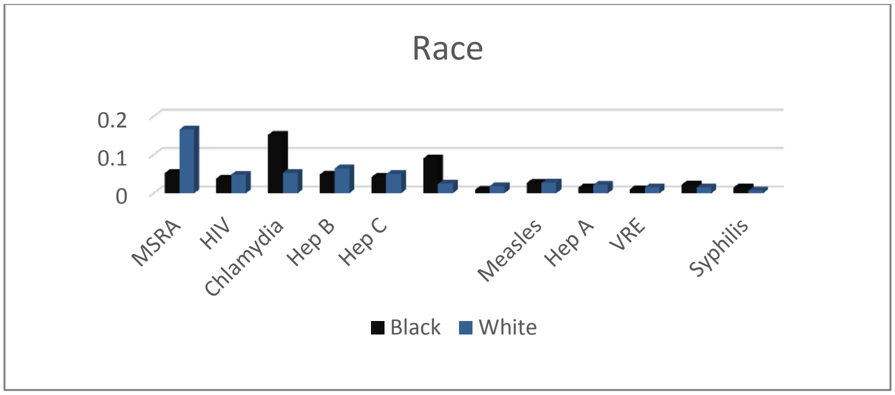


Figure 5. Diseases Distribution Across Race

Query efficiency

We test the database efficiency by three sets of common samples queries designed by visualization and health science experts. The first query set is about geographical distribution of one or a combination of diseases. The second query set retrieves strong associated diseases to a given disease. The third query set finds common diseases occurring at a given range of age. Table 4 summarizes the performance of three types of queries and suggests that the database is optimized and effectively support real time association mining.. The performance of aggregated queries is satisfactory and may be further optimized. .

Table 4: Three sets of query used for testing the database

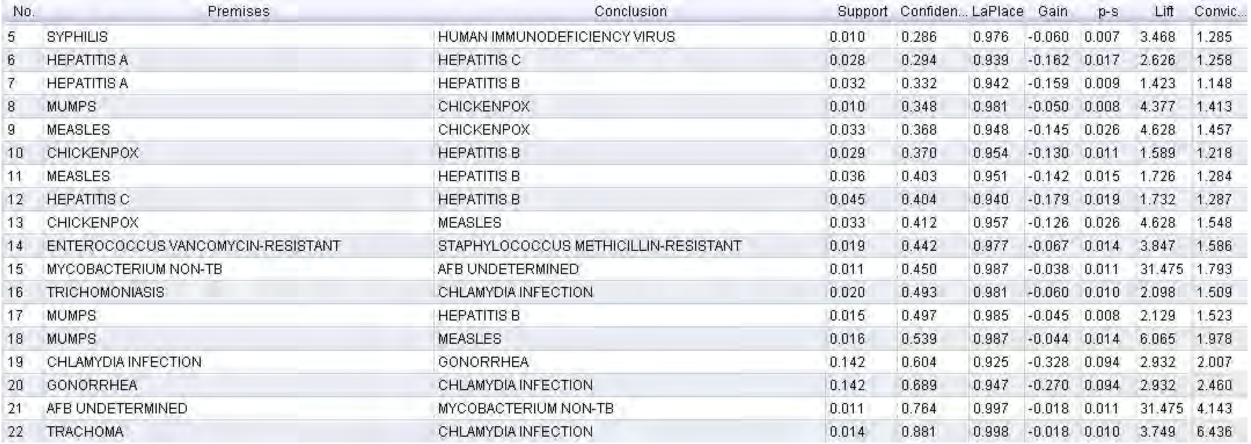
Query set	Example	Involved tables	Runtime
1	Geographical (at city level) distribution of chlamydia	location, diseases, patient-disease occurrences	12s
2	List the diseases associating with chlamydia	diseases, associations	0.5s
3	What are the most common diseases for patient age from 20 to 40	diseases, patients, patient-disease occurrences	16s

Association Mining

The database contains significant associations which are not widely reported in literature, such as Antidarrheal treatment and runny nose symptom (confidence: 0.73), screla and Tylenol treatment (confidence 0.70), posturing and Mortin treatment (confidence: 0.81), etc.

Table 5 shows part of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the support, confidence, Laplace, Gain, p-s, lift and Conviction. We are working with domain experts on evaluating the association rules and tuning the parameters to produce optimum result.

Table 5: Association Rules among Diseases



No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convic...
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.626	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.368	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.954	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.987	-0.038	0.011	31.475	1.793
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.129	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.460
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Clustering Analysis

We found several cluster containing close relationships between diseases and terminologies, such as {Biliary Sludge, HFA, Macrocytosis, Paroxysmal, Pseudogout, back discomfort, betimol, hesitancy, Intron A}, {Gastric Polyps, Kidney failure, antral, benefix, benicar}, {Duodenal Ulcer, Helicobacter Pylori, Malabsorpition, amylase, antimetics} and {appetite lost, immunoglobulin, retrovir} , etc. Some clusters highly correspond to specific diseases or medical processes. For example, appetite lost, immunoglobulin, retrovir are HIV related symptoms. Meanwhile, some clusters contain diseases and terms associated with several medical processes. For example, we found a cluster including gastrointestinal terms (Duodenal ulcer, Helicobacter pylori, Malabsorption, acyclovir, amylase and antiemetics), additive behavior (drinking, lortab andmarijuana) and cancer (methotrexate, vincristine and zofran). This cluster may suggest negative impact of addictive behavior toward digestive system. The appearance of cancer drugs in this cluster could raise a research question about the impact of additive behavior toward the metabolism process, which will further affect the cancer drug efficiency.

Sequential Patterns

We found 105 disease-associations satisfying all 3 criteria about rule_confidence, begin_before_end and coverage to construct the frequent sequential disease patterns. We found 3 groups of sequences in the NCD data. The first group contains only one sequence Hyperplenism □ Annemia. The second group contains 5 diseases: Fibrosis Pulmonary, Staphylococcus Methicillin-resistant, Biliary Stricture, Cycsticercosis and Meconium Ileus, in which Fibrosis pulmonary and Meconium ileus stay at the triggering position. This disease group may raise additional research questions since these diseases occur at different organs. The last group is marked by Hepatitis A and 56 other diseases staying at the triggering position of Hepatitis A.

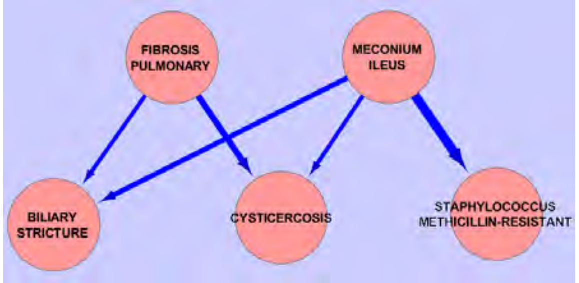
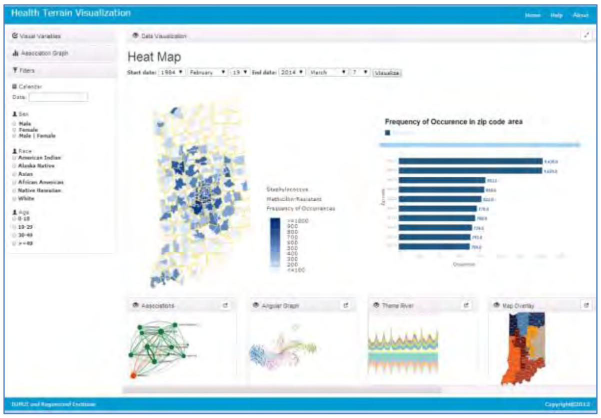


Figure 6. Fibrosis/Meconium-ileus sequence

Visualization

We have developed a Health Terrain Visualization system to visualize interesting knowledge associations that emanate from the data mining and text mining processes. The visualization environment provides the user with a variety of ways to observe interesting patterns. For example, Figure 7a shows the map of Indiana, structured by county and zipcode, displaying the occurrence of the disease *Staphylococcus Methicillin-resistant* across the state as a heat map. The user will have the ability to use filters to observe patterns of various diseases in a similar manner.



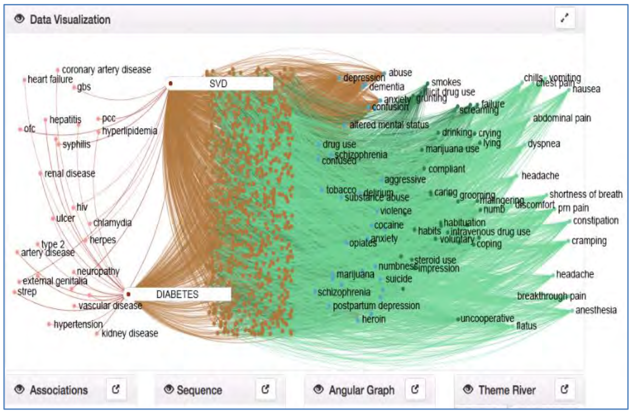


Figure 7a. Heat map of disease

Figure 7b. A multi-feature association graph

In Figure 7b, first loose column of peach colored nodes on the left are the *comorbidities*. Then the *diseases* are highlighted with maroon nodes. The next columns of brown nodes are *patients* followed by the light blue nodes representing *mental behaviors*. The dark green nodes are *risky behaviors*, and the final columns of light green nodes are *symptoms*.

Discussion

NLP provides a means to increase the amount of information that is present in the NCD data. Further analysis is being carried out based on location information that is present in the NCD dataset. This kind of location-based-mining is very useful for public health practitioners in understanding the nature of a disease. To achieve this, we are mining the NCD data and the corresponding clinical data simultaneously, based on zip codes, counties and cities. This work is being incorporated into our Health terrain visualization system.

In this study, by finding associations and grouping strongly-related terminologies into clusters, data mining is useful in guiding visualization application to adopt better visualization layout and highlighting significant and useful information to the users. We also design a visualization-oriented database model to reduce the heavy data-fetching and computation workload on the visualization application.

In addition, we prove that data mining could discover unreported/ill-reported associations between various terminologies in health data science, such as medication-symptom, disease comorbidity, etc. We are working with health professionals to validate and explain the new-found associations. On the other hand, this fact may open further collaboration between computational approaches and traditional biological-medical ontology approach to achieve better understanding on the mechanism of the development and spread of diseases. Another promising direction is integrating EHR data mining with genotype information to construct in-depth knowledge about disease and drug mechanism and visualize this integration by GeneTerrain[8].

Acknowledgement

The project is supported by Department of the Defense; award number W81CWH-13-1-0020. We gratefully acknowledge the contribution of Jennifer Williams; Regenstrief Institute, Project Manager.

References

1. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
2. Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html
3. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11
4. Chapman , W.; Bridewell , ; Hanbury , ; Cooper , G. F.; Buchanan , G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 2001, 34 (5), 301–310.
5. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
6. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342
7. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithm. Massachusetts Institute of Technology Press, 2009, pp 390-396
8. You Qian, Shiaofen Fang, and Jake Y. Chen. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks (2008) Information Visualization, doi: 10.1057/palgrave.ivs.9500169. manuscript.
9. Mohamemed J. Zaki and Wagner Meria. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press 2014, Great Britain. pp 336-337
10. Van Mechelen I, Bock HH, De Boeck P. Two-mode clustering methods:a structured overview. Statistical Methods in Medical Research 13 (5): 363–94, 2004

Data Exploration of a Notifiable Condition Detector System

Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[1] , Sam Bloomquist[1], Anand Krishnan[2], Jeremy Keiper[3], Shaun Grannis, MD[3]
1 Department of Computer Science, Indiana University – Purdue University, Indianapolis;
2 School of Informatics and Computing, Indiana University – Purdue University, Indianapolis;
3 Regenstrief Institute, Indianapolis, IN

Abstract

The Notifiable Condition Detector (NCD) system is an automated electronic lab reporting (ELR) and case-notification system developed by Regenstrief Institute. It has been used in Indiana for over ten years to report laboratory results for the detection of notifiable conditions such as novel H1N1 influenza, sexually transmitted diseases, lead poisoning, and salmonella [1]. In this paper, we discuss ongoing efforts to analyze and visualize dimensions of the NCD data. We identify most common conditions, describe the distribution of the diseases across gender and race, study the co-occurrence of diseases and find the association rules among different diseases.

Introduction

The Regenstrief Institute implemented and has maintained an HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in Indiana. The Notifiable Condition Detector (NCD) uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC). The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations [1]. The NCD automatically detects positive cases of pre-specified conditions and forwards alerts to local and state health departments for review and possible follow up. These alerts enable public health to conduct more effective and efficient population health monitoring.

The initial analytic dataset contained 833,710 notifiable cases from 543,209 distinct patients. The dataset was deidentified by removing HIPAA identifiers and replacing patient age and zip code columns with pseudonymized values. The dataset contains 22 columns in total, including patient pseudo ID, condition name, test result name, test result value, test normal range, race, and gender, among others. The missing data rate for columns varied substantially from 0% for column Patient Pseudo_ID to over 70% for column Test_Abnormal_Flag.

Data Analysis

We analyzed the condition distribution across genders and and the result is shown in figure 1.

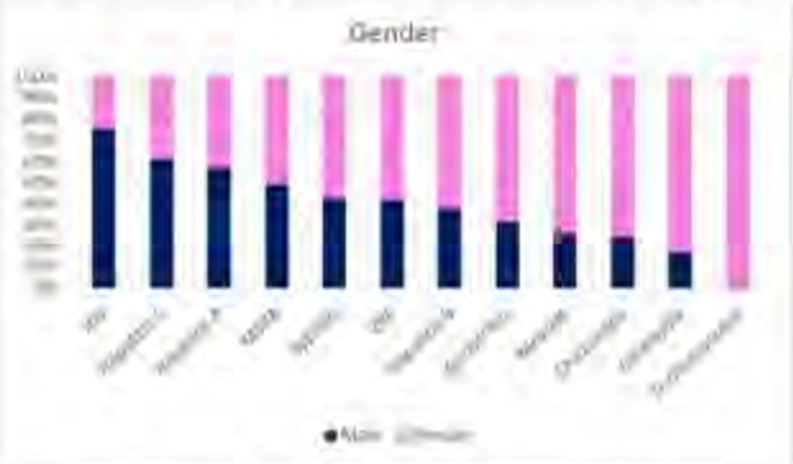


Figure 1: Condition distribution across gender

The conditions in the figure are listed according to male proportion in descending order. It shows that HIV and Hepatitis C are more common in men than in women, while Trichmoniasis , Chlamydia, Chickenpox, Measles, Gonorrhea, and Hepatitis B are more common in women than in men. The conditions that do not show significant gender difference include Hepatitis A, MRSA, Syphilis and VRE.

Disease Association Mining

In the NCD dataset, there are 39,771 patients with two or more diseases. We studied the co-occurrence pattern and derived the association rules between diseases. We generated a graph of associations among conditions using RapidMiner version 5.3, illustrated in Figure 2. It shows 2 clusters among the conditions. One cluster includes Chlamydia/Trachoma, Gonorrhea and Trichomoniasis. The other larger cluster includes Hepatitis A, Hepatitis B, Hepatitis C, HIV, Syphilis, Measles, Mumps, and Chickenpox. The figure also suggests an association between MRSA and VRE. The reason behind the associations has yet to be studied with domain experts.

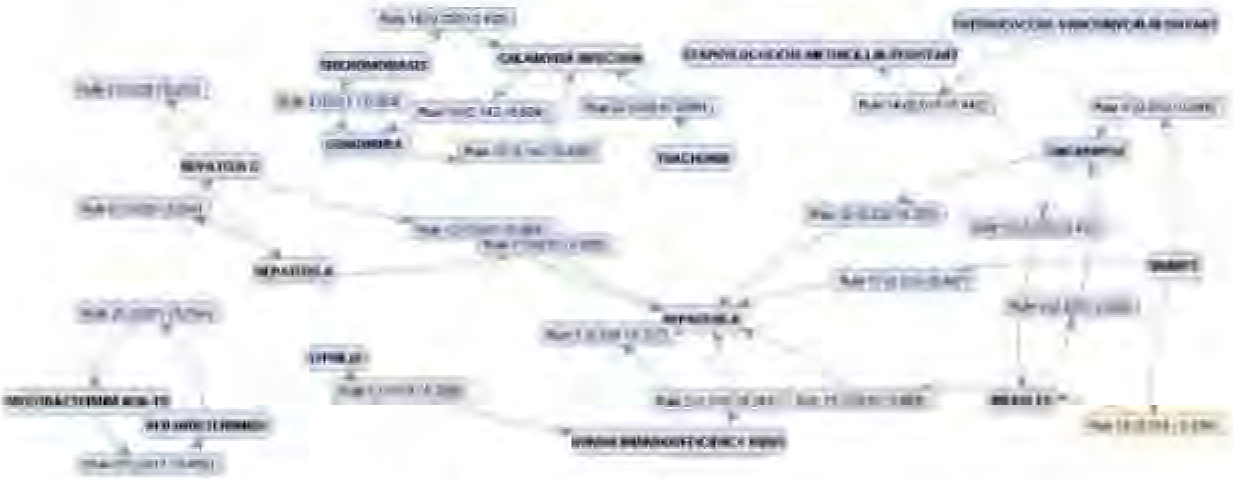
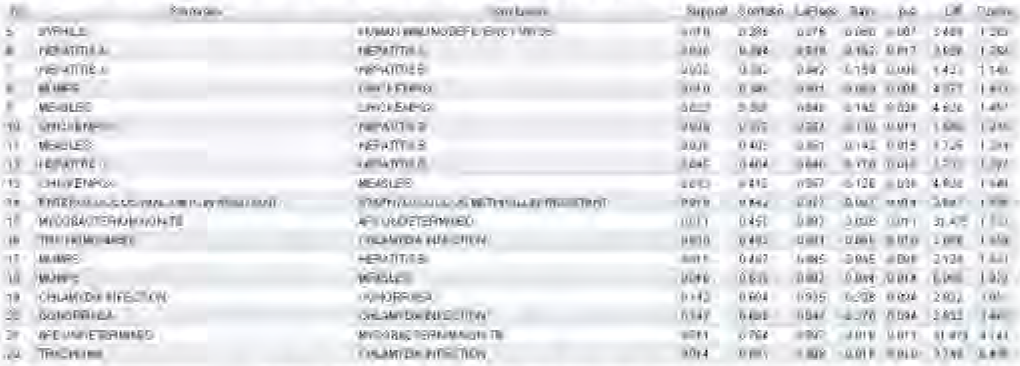


Figure 2: Diseases Associations

Table 1 shows a portion of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the Support, Confidence, Laplace, Gain, p-s, Lift and Conviction. The rules are sorted according to the Confidence in ascending order. The higher the Confidence, the stronger the association is. A complete explanation of all the quality measures can be found in [2]. We are working with domain experts to evaluate the association rules and tuning the parameters to produce optimum results.



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.018	0.285	0.976	0.080	0.007	2.489	1.203
6	HEPATITIS A	HEPATITIS C	0.030	0.394	0.916	-0.162	0.017	3.609	1.264
7	HEPATITIS A	HEPATITIS B	0.032	0.382	0.962	-0.159	0.006	1.421	1.140
8	MUMPS	CHICKENPOX	0.010	0.386	0.991	-0.069	0.008	4.077	1.400
9	MEASLES	CHICKENPOX	0.027	0.390	0.940	-0.145	0.020	4.626	1.491
10	CHICKENPOX	HEPATITIS B	0.026	0.375	0.952	-0.130	0.011	1.696	1.246
11	MEASLES	HEPATITIS B	0.026	0.402	0.951	-0.142	0.015	1.726	1.284
13	HEPATITIS C	HEPATITIS B	0.045	0.404	0.946	-0.176	0.048	1.710	1.297
15	CHICKENPOX	MEASLES	0.027	0.412	0.957	-0.126	0.020	4.626	1.540
16	ENTEROCOCCUS VANCOMYCIN RESISTANT	STAPHYLOCOCCUS METHICILLIN RESISTANT	0.019	0.442	0.957	-0.042	0.014	3.887	1.590
17	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.451	0.993	-0.026	0.011	31.475	1.770
18	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.010	0.483	0.981	-0.061	0.010	1.998	1.458
17	MUMPS	HEPATITIS B	0.011	0.487	0.985	-0.045	0.006	2.124	1.441
18	MUMPS	MEASLES	0.016	0.616	0.982	-0.044	0.014	6.466	1.819
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.935	-0.038	0.024	2.602	1.941
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.606	0.947	-0.376	0.094	2.652	1.440
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.141
24	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.991	0.999	-0.014	0.010	3.746	8.409

Table 1: Association rules among conditions

Acknowledgement

The project is supported by Department of the Army; award number W81CWH-13-1-0020.

References

1. Automated Electronic Lab Reporting and Case Notification, http://www.regenstrief.org/cbmi/areas-excellence/public-health/.

Detecting Comorbidity of Chlamydia from Clinical Reports

Mathew Palakal, PhD[1], Shiaofen Fang, PhD[2], Yuni Xia, PhD[2] , Shaun Grannis, MD[3], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[2] , Sam Bloomquist[2], Anand Krishnan[1], Jeremy Keiper[3]
[1]School of Informatics & Computing, Indiana University – Purdue University, Indianapolis;
[2]Department of Computer Science, Indiana University – Purdue University, Indianapolis;
[3]Regenstrief Institute, Indianapolis, IN

Abstract

Using a standards-based messaging and vocabulary infrastructure, the Regenstrief Institute implemented and has maintained an unparalleled automated electronic laboratory reporting and noticeable condition detection (NCD) system for over 11 years [1]. The NCD automatically detects positive cases of pre-specified conditions and forwards alerts to local and state health departments for review and possible follow up. In this paper, we discuss ongoing efforts to analyze the clinical reports of one specific NCD condition, Chlamydia. Our goal is to identify the presence of any comorbidities of Chlamydia across 6238 patient records and integrate this finding along with our health analytics and visualization system that we are developing.

Introduction

The participants of this project are working on a Health Terrain visualization system centered on an innovative concept-based knowledge discovery and visualization. In this approach, raw health data are first processed, mined and transformed to an information-rich Health Concept Space, where attribute values, association relationships and other partial knowledge are extracted from patient data to form a more structured multi-dimensional data space. The concept space is built on a controlled vocabulary (concepts) that can be pre-defined based on application needs, and therefore is a scalable framework that can be expanded progressively as the applications and use cases expand. For building the concept space, we utilize the structured data as well as unstructured clinical reports from the EHR data. One piece of knowledge that can be extracted from the clinical notes is comorbidity and this can be discovered using Natural language processing (NLP). The text in medical records (e.g. radiology reports, pathology reports, clinical notes, and discharge summaries) includes a wealth of information about patients. NLP can be very useful in extracting information from these free text documents and creating structured information that can be used for further knowledge extraction.

Data

The dataset consists of 6238 de-identified clinical notes that include discharge summary, laboratory reports, etc. Since the clinical records are de-identified, the patient specific information is lacking in the clinical reports.

Method

Natural language processing (NLP) techniques are carried to detect comorbidities that co-occur with chlamydia. In this case, the NLP process is composed of low-level and high level task. Low level tasks consist of sentence splitting, tokenization, stemming, part of speech (POS) tagging and phrase chunking (identifying phrases from POS tagged tokens) and the higher level tasks consists of named entity recognition (NER) of co-occurring diseases. Once the NER process is complete, we use the *tf-idf* vector space model [1] (term frequency * inverse document frequency model) to identify significant co-occurring diseases along with chlamydia. The *tf-idf* is recognized as one of the more effective text mining models compared to Log Level Likelihood and Odds ratio models, amongst the various other models. The *tf-idf* model uses the concept of relevance and co-occurrence of terms. The relevance of a term *j* w.r.t. a document *i* is given as,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j}) \tag{1}$$

w_{ij} = relevance of term j in the patient record i; T_{ij} = term frequency of term j in in the patient record i; N_j = frequency of records for term j; N = total number of records (N=6238). A particular term is more relevant w.r.t. a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight is attached with every association between a pair of terms [3]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k}) \tag{2}$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. If the terms do not co-occur in any of the *N* records, then the association is 0. We will be interested mainly in non-zero associations.

Results

After applying the low-level NLP steps, the resulting terms were subjected to named entity recognition (NER). The UMLS [4] database was used for NER to identify diseases that are present in the 6238 discharge summaries. A total of 1337 possible disease conditions were identified that can be roughly considered as comorbidities with Chlamydia. Figure 1 show the most commonly occurring diseases along with the number of reports in which they occur. The co-occurring diseases with Chlamydia were further analyzed using the *tf-idf* model to understand the significance of these diseases across all the 6238 records. For this analysis, *tf-idf* score was calculated for the diseases using the Equation 1.

Most Occurring Diseases in # of Reports

Figure 1: Top diseases with the corresponding number of reports

Comorbidity with Chlamydia

The final analysis and the goal of this work are to identify comorbidity with Chlamydia. To accomplish this, we calculated the pair-wise significance of each disease with Chlamydia using Equation 2. Figure 2 shows all the diseases those are found to be comorbid with Chlamydia. This graph shows some of the expected diseases such as hepatitis, gonorrhea, etc., along few other health conditions such as Bacterial vaginosis, obesity, and so on. The accuracy of these findings has to be validated by the experts. In addition to the comorbidity, the NLP analysis also revealed symptoms, risky behaviors, and mental behaviors those are associated with chlamydia as shown in Figure 3.

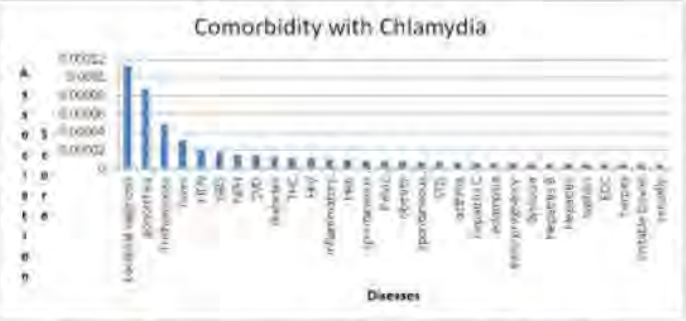


Figure 2: Top scoring comorbidity with Chlamydia

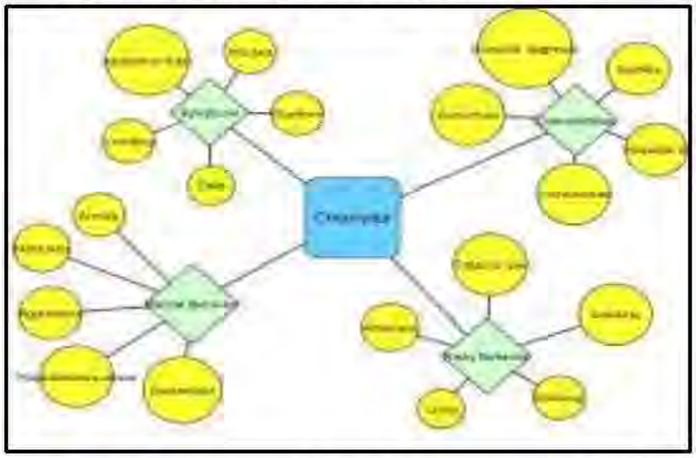


Figure 3: Symptoms and risky behaviors associated with Chlamydia detected using NLP

Conclusions

We have developed an NLP method to identify comorbidity of Chlamydia from the discharge summary of 6238 clinical records. This information, once validated by the experts, will be incorporated in to our Health Terrain visualization system that is currently under development.

Acknowledgement

The project is supported by Department of the Army, award number W81CWH-13-1-0020.

References

1. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
2. G. Salton. Introduction to modern information retrieval. McGraw-Hill, New York, 1983
3. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342
4. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–1

Use Cases for Public Health Data Visualization

Jeremy Keiper[1], Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], Shaun Grannis, MD[3], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[1], Sam Bloomquist[1], Anand Krishnan[2]
1 Department of Computer Science, Indiana University – Purdue University, Indianapolis;
2 School of Informatics, Indiana University – Purdue University, Indianapolis;
3 Regenstrief Institute, Indianapolis, IN

Abstract

Epidemiologists in Marion County, Indiana, use massive amounts of public health data to provide insight to previous and current disease outbreaks. Typically, they use statistical analysis and simple charts to convey information in public reports, but the work is tedious and difficult without contextual knowledge from years of investigation. Use cases highlighting potential disease outbreaks and research topics, provided by these epidemiologists, will inform a new interactive health data visualization system designed to facilitate and enhance the discovery and strategic intervention.

Introduction

Epidemiologists and other public health officials need to contain the spread, and identify and reduce exposure to disease. We intend to provide a visualization engine to facilitate discovery of these outbreaks. The following factors contribute to a useful public health dataset, and can be used as variables in visualizations:

- Age
- Gender
- Geographic Location
- Race
- Social Networks

Existing visualizations are limited to heat maps or simple overlays, typically simplified to charts and tables for communicating the most essential data to describe the spread or impact of a disease. A useful interface will allow for rendering these reduced views for use in reports. Data analytics can provide insight to what might be the most interesting data, with the final decision of relevance in the hands of the user.

Interviews with Dr. Joseph Gibson, MPH, PhD, Director of Epidemiology at the Center for Urban Health in the Marion County Public Health Department, and Dr. Roland Gamache, PhD, Affiliate Research Scientist with Regenstrief Institute, inform the use cases in this presentation. Some use cases are specific to Indiana or the Indianapolis area.

Use Case: Histoplasmosis

A rare disease, affecting only 20-25 people yearly in Marion county, histoplasmosis outbreaks should be easy to detect locally. Excavation that disturbs ground where mold spores have settled typically predicates an outbreak of this disease. Indiana has a long history with histoplasmosis due to a high concentration of relevant spores throughout the state; the largest outbreak in the country occurred in Indianapolis in September 1978 and again in August 1979 [1]. Public health officials need to see each diagnosis, with multiple occurrences in a similar location prioritized over others, alongside relevant excavation data. Visualization techniques can highlight geographic proximity to potential causes.

Major Visualization Factors	Visualization Traversal Vectors
Occurrence density in geographic locations	Chronological time
Nearby potential causes (e.g. construction, tree removal, or demolition)	Proximity to causal events

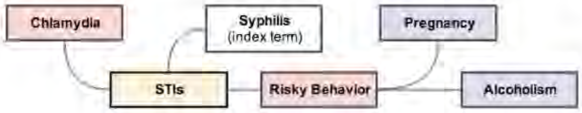
Use Case: Gonorrhea

When researching gonorrhea, epidemiologists are interested in co-morbities and other related diseases. They also want to know who is repeatedly infected after treatment. Gonorrhea is a social disease, transmitted sexually, and a proper visualization will highlight these connections. Social networks, both online (e.g. Facebook and Twitter) and traditional (e.g. high schools, colleges, neighborhoods, and workplaces), become the most important correlation factor. Intervening with super-spreaders, people who may be unknowingly giving gonorrhea or related diseases to several others, can help contain the impact and prevent recurrence. Social data may require approval by individuals, but would facilitate better care.

Major Visualization Factors	Visualization Traversal Vectors
Frequency of recurrence in individuals	Chronological time
Comorbidities (e.g. other sexually transmitted diseases)	Social distance from a super-spreader

Use Case: Risky Health Behavior

Some diseases are indicative of risky health behavior, and epidemiologists need to know when an individual has a higher potential for infecting others in the community. Visualizing layers of an index term against related diseases and risky behavior can show unexpected correlations. The example below shows first, second, and third relations.



Major Visualization Factors	Visualization Traversal Vectors
Cumulative occurrences of risky behavior	Time deltas (e.g. daily, one week, two weeks, months)
Time range between occurrence (e.g. a spike on multiple short deltas)	Age ranges Relatedness of other diseases or categories

Use Case: Influenza Pandemic

Influenza is the type of disease that appears seasonally, as a response to airborne allergens. Strategies for a breakout of influenza vary depending on the current state of the pandemic. When integrated with a monitoring system, a visualization engine could highlight a grouping of recent diagnoses and assist in intervention and containment. If the pandemic has already breached a certain threshold, the system would switch to a different visualization useful for monitoring and forecasting spread. It is important in such a breakout to identify the first cases, to understand how the disease is progressing and spreading from the index patients. This information should be presented visually to help decide the best strategy.

Major Visualization Factors	Visualization Traversal Vectors
Earlier occurrences highlighted for easy identification	Amount of time since infection
Occurrence density in geographic locations	Social or geographic location links

Future Research

Each use case above demonstrates problems in public healthcare that cannot be easily interpreted without a series of statistical charts and tables, and the knowledge to recognize trends and similarities. We will use these scenarios as starting points to inform the design of a visualization engine with an interactive interface, primarily useful to the epidemiologists. We expect this system to make the discovery and research of outbreaks and pandemics easier, faster, and more effective.

Acknowledgements

The project is supported by the Department of the Army, award number W81CWH-13-1-0020

Visualizing Large Healthcare Data by Geospatial Texturization

Shiaofen Fang, Shenhui Jiang, Sam Bloomquist, Mathew Palakal, Yuni Xia, Li Huang, Jeremy Keiper, and Shaun Grannis

ABSTRACT

Healthcare data visualization is challenging due to not only its size and complexity, but also the needs for integrating geospatial information, temporal information, and multi-dimensional attributes within a common visual context. In this paper, we present a new visualization technique called geospatial texturization, suitable for applications involving large healthcare data. In this approach, multi-dimensional and time-varying data can be encoded in a texture space, and then mapped to the geospatial surfaces. Two texturization techniques are developed: (1) a noise texture pattern using a turbulence function is constructed to embed multiple attributes within geospatial regions; (2) a new offset contouring method is proposed to represent time-varying data or multi-attribute data using boundary contours of the geographical regions. A radial coordinate based Ring Graph technique is also developed as a supplementary visual representation for more detailed patient-level data. This work is implemented under a general healthcare data visualization framework called HealthTerrain.

Keywords: spatiotemporal visualization, geospatial information visualization, healthcare data, visual data mining.

1 INTRODUCTION

As electronic healthcare systems are being fully integrated nationally, the effective visualization of large and complex healthcare data becomes increasingly desirable for timely decision making and trend/pattern detection [1]. The problem, however, is very challenging for several reasons:

1) Health data is a data-rich, information-poor domain. In Electronic Health Record (EHR) systems, data are almost always heterogeneous, unstructured, hierarchical, and longitudinal.
2) EHR systems are often extremely large. While it is possible to visualize an EHR system in small scales and with a focused scope, high impact knowledge discoveries more likely come from global scale (population-wide) visualization and knowledge mining.
3) Visualizing population-level health data often involves presenting geospatial and time-series data in a common visual context. This presents a challenge in visual encoding of the information space.

For heterogeneous and complex data, feature extraction through data mining is critical, as visualizing a feature space is much more feasible. For healthcare data, this feature space often consists of healthcare terms (ontology) and their relationships. Therefore, the effective integration of data processing, data mining, and text mining is necessary in healthcare data visualization. Although healthcare data is very large, the visualization of aggregated features, combined with some patient level visualization, can be very effective in revealing the patterns and trends of population health. It is therefore important to develop multiple visualization methods that can show different perspectives of the data.

One of the unique challenges in healthcare data visualization is how to visualize multi-attributes and time-series data with associated geospatial information. In our approach, we embed multiple attributes and the time variable within a geospatial representation to take advantage of the available geographic space. This can be done by mapping texture images onto the geospatial surfaces. The key is then to properly represent the multi-attributes and time-series information in a texture image by constructing visually effective texture representtaions. We call this process texturization.

In the rest of this paper, related work will be discussed in Section 2. In Section 3, we will ouline the high level features of a healthcare data visualization system we have developed, as well as the data mining and text mining techniques used in this system. Section 4 will focus on the visualization techniques including the two texturization methods and the Ring Graph method. Conclusions and future work will be discussed in Section 5.

2 RELATED WORK

The visualization of large scale healthcare data has not been extensively studied. There are several existing works and visualization systems that deal with the secondary use of electronic health record data in a limited scope. LifeLines [2] uses a traditional 2D time line visualization technique to visualize specific patient medical and health history. It emphasizes the visualization of temporal ordering of events with limited aggregation effect. An extension of LifeLine, LifeLine2 [3], enables multiple patient comparisons and aggregation for analysis, but the visualization design limited its scalability. A similar system, call TimeLine [4], re-organizes and re-groups multiple EHR content types in a layout of Y-axis to track multiple events along the same time line. In [5], a set of visualization tools are described for visualizing a patient's electronic health record to aid physicians' diagnosis and decision-making. The traditional matrix view and parallel coordinates are the main techniques applied. CLEF [6] is a system enabling visual navigation through a patient's medical record using semantically and temporally organized networks to represent events throughout the patient's medical history. CLEF also supports limited text processing capabilities for generating textual summaries. None of these existing systems is capable of visualizing large-scale integrated EHR datasets. A review paper on visualization tools for infectious diseases is given in [7].

The geospatial visualization of time-series data is challenging because it is difficult to encode the time axis in a geospatial context. Animation based techniques (e.g. [8]) do not provide a good space-time overview. Other techniques, such as color-coding of time [9], connecting time-lines [10], and time-curves [11], often introduces visual clutter and occlusion, which are infeasible for large scale datasets. A well-known technique in geospatial time-series visualization is Space-Time-Cube [12-16]. It is a 3D representation of a combination of time axis (Z-axis) and a 2D geographic map (X-Y plane). Time-lines or time-curves are used to depict data evolution over time. While time and spatial information are integrated in a 3D visual representation in a

*

space-time-cube, the sense of space-time embedding diminishes as the data moves up in the time axis. Visual clutter will also be a problem with large datasets.

Texture-based visualization techniques have been widely used for vector field data, in particular, flow visualization. Typically, a gray-scale texture is smeared in the direction of the vector field by a convolution filter, for example, the Line Integral Convolution (LIC), such that the texture reflects the properties of the vector field [17,18,19]. Similar techniques have also been applied to tensor fields [20,21]. Another class of related work is pixel-oriented techniques [22], which map each data value to a colored pixel and present the data values belonging to one dimension (attribute) in a separate subwindow. Examples include spiral [23], recursive pattern [24], circle segment techniques [25], and hierarchical temporal patterns [26]. These texture-based or pixel-oriented techniques are fundamentally different from our texturization techniques since in texturization, textures are not used to represent the individual data points, but to depict the overall trend and strength of the data in geospatial regions.

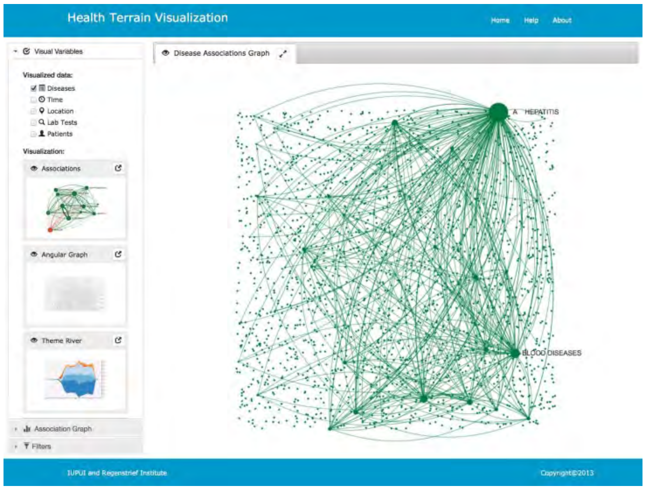
3 The HealthTerrain System

The visualization techniques to be presented in this paper is part of a prototype visualization system we are developing called HealthTerrain, It is a browser-based interactive visualization system for large healthcare data. It employs a concept space approach that combines data mining, data processing and text mining techniques to effectively transform the heterogeneous, unstructured, hierarchical, and longitudinal data to a uniform space of a controlled ontology and their associations. Multiple visualization toolkits with data filters are designed to visualize different scales of the data. The combination of details patient level visualization and aggregated data features is an effective way to visualize large scale data sets.

To test our HealthTerrain visualization system we used a large public health notifiable disease reporting system. The Regenstrief Institute implemented and maintains an unparalleled HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in the State of Indiana. The Notifiable Condition Detector (NCD) System uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC) [27]. The NCD dataset contains 833,710 public health notifiable cases spanning more than 10 years from among 439,547 unique patients. An additional dataset containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories were extracted from the Indiana Network for Patient Care (INPC) health information exchange for these patients. In order to comply with the patient privacy policies and protocols of the institutes where the datasets came from, the actual data visualized in this paper has been altered or perturbed. Nevertheless, the overall patterns and trends of the data are still preserved.

The "concept space" represents a uniform layer of clinical observations and their associations, and enables users to explore data using various visualization and analysis methods. Concept terms are derived from data mining and text-mining processes applied to the use case datasets. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including disease, symptom, mental behavior, and risky behavior. We also performed additional analysis to compute the correlations among conditions using the tf-idf (term frequency – inverse document frequency) vector space model to identify the significantly co-occurring diseases

An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology. Figure 1 shows the associate map within the system's web-based interface.



☑ Diseases
□ Time
□ Location
□ Lab Tests
□ Patients

Figure 1: Disease association map and the web interface of the HealthTerrain system.

4 Visualization Algorithms

Association map (Figure 1) is a graph visualization of the association relationships among the diseases and other terms in the concept space. It can serve as a platform supporting interactive selection of concepts to dynamically visualize data using a variety of tools in the visualization system. To draw an association graph, a spring-embedder algorithm [28] is used to layout the graph nodes. Nodes picked on the association map are then be visualized with geospatial information, possibly with time varying variables. In the rest of this section, we will first describe the two types of texture generation techniques, and then discuss the terrain surface method. Finally we will describe the Ring Graph visualization method for patient level visualization.

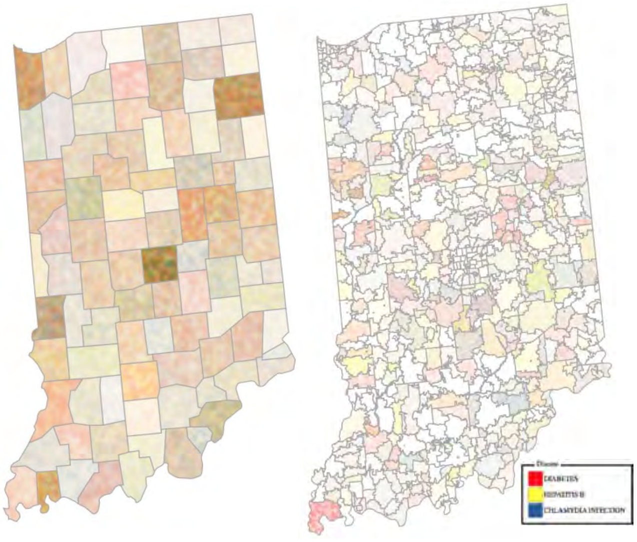
4.1 Texturization

In texturization, texture images are constructed to represent the overall data trends and distributions in different geospatial regions. Once the textures are generated, we will first visualize them on a 2D geographic map as a heatmap image, and then map them to terrain surfaces. There are two different types of textures that will be generated here (1) noise pattern texture for the representation of multiple attributes; and (2) offset contour texture for the time-varying data representation.

4.1.1 Noise Texture

We aim to represent multiple attributes for each geographic region using color coded texture patterns so that the users can easily perceive the representations of different attributes, not only within one region, but also its overall geospatial distributions across many regions in a geographic area (e.g. a state).

We first construct noise patterns to create a random variation in color intensity, similar to the approach in [29]. Different color hues will be used to represent different types of attributes, for example the occurrences of different diseases. A turbulence function [30] will be used to generate the noise patterns of different frequencies (sizes of the sub-regions of the noise pattern). These multi-scale patterns may be applied to different scales of geographic areas (e.g. counties vs zip-codes). Since the noise pattern involves the mixing and blending of different color hues, we choose to use an RYB color model instead of RGB model, as proposed in [29], since RYB color model provides more intuitive representation of the weights of different colors after blending. Figure 2 shows two examples of the heatmap views of three diseases, Diabetes. Hepatitis B, and Chlamydia, over the Indiana state map.



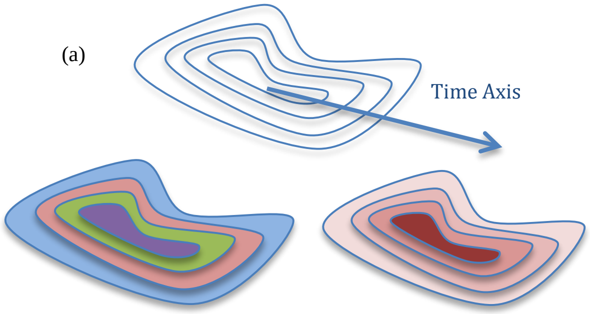
(a) (b)

Figure 2: Heatmap views of noise textures over the Indiana state map: (a) county based; (b) zip-code based

4.1.2 Offset Contouring

Offset contouring is designed to represent attribute changes over time within a geographic region. It can also be used to represent multiple attributes. Similar to the Noise Pattern approach, we first construct a texture image using offset contour curves to form shape-preserving sub-regions. We will then use varying color shades or hues to fill the sequence of sub-regions to represent the change of attribute values over time, or to simply fill the sub-regions with different color values to represent multiple attributes.

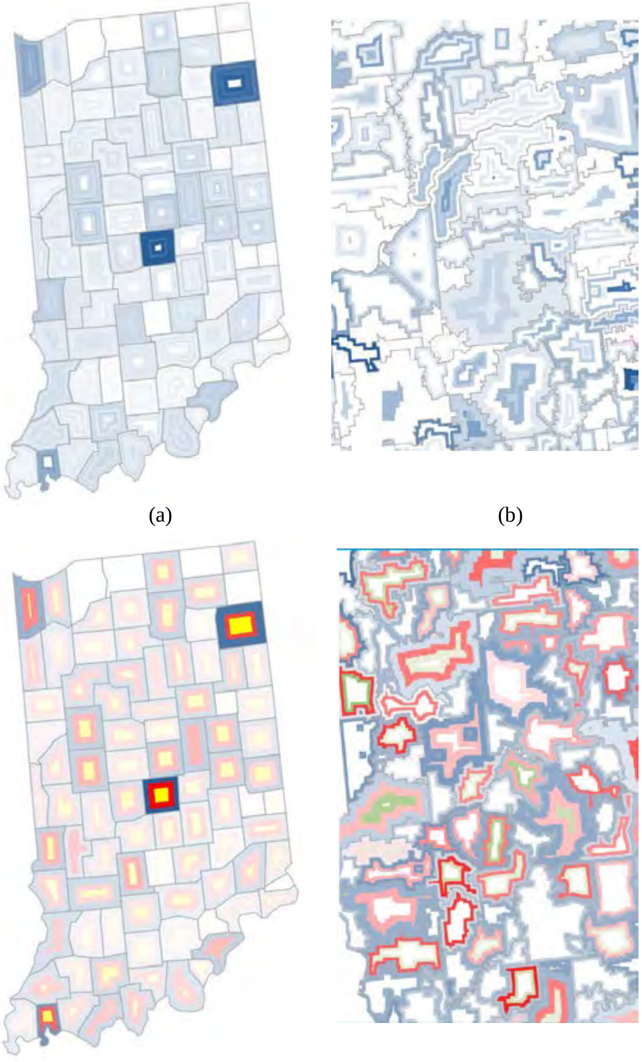
The offset contours are generated by offsetting the boundary curve toward the interior of the region, creating multiple offset boundary curves (Figure 3a). There are several offset curve algorithms available in curve/surface modeling. But since in our application, the offset curves do not need to be very accurate, we opt to use a simple image erosion algorithm [31] directly on the 2D image of the map to generate the offset contours. Figure 3b and 3c shows the color-filled sub-regions after offset contouring.



(b) (c)

Figure 3: Offset contouring. (a) Offset contours; (b) Multi-attribute coloring; (c) Time-series coloring

In time-series data visualization, the time line can be divided into multiple time intervals and represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time (Figure 3c). This approach, however, has two limitations. First, when the boundary shape of a region is highly concave, the image erosion technique sometimes does not generate clean offset contours. This usually can be corrected using a geometric offset curve algorithm such as the one in [32]. A second limitation of this approach is that it requires a certain amount of spatial area to layout the contours and color patterns. In public health data, however, these attributes are typically defined on geographic areas, which provides a perfect platform for texturization. Figure 4 shows a few examples of the heatmap views of offset contouring over the Indiana state map. Figure 4 (a) and (b) show the time-series views of Influenza, from 2004 to 2012. The time interval is divided into 8 subintervals. Figure 4 (c) and (d) show three diseases, Influenza, Typhoid Fever, and Hepatitis B.



(a) (b)

(c) (d)

Figure 4. Heatmap views of offset contouring over the Indiana state map: (a) County based time-series data; (b) Zip-code based time-series data; (c) County based multi-diseases data; (d) Zip-code based multi-diseases data.

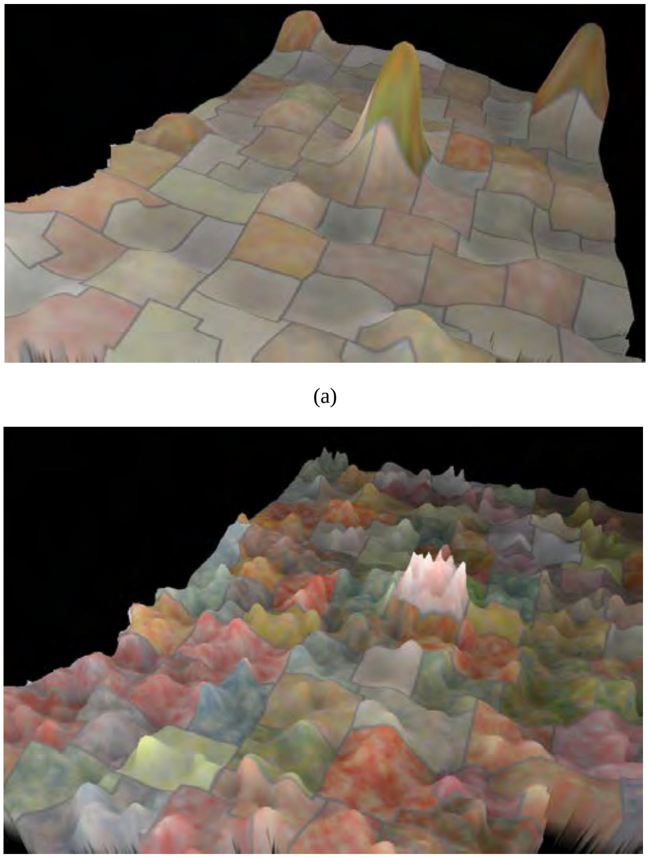
4.2 Texturized Terrain Surface

The heatmap views are effective in conveying the relative distributions of multiple attributes in different regions of a geographic area. The distribution of the total attribute values, however, becomes more difficult to perceive as the information has been disbursed by the texture patterns. This problem can be resolved by mapping the texture pattern onto a 3D terrain surface using the total attribute values as a height field.

A 3D surface can be constructed on top of a geographical region (e.g. the map of Indiana State). Typically, data are aggregated to individual geographical regions, such as counties and zip-codes, to form a height field. The height value can be, for example, the sum of multiple attribute values in a region, or the total occurrence of an attribute over the given time period for time-series data. To construct the surface, 3D scattered interpolation technique is applied so that every pixel point within the geographical boundary will have an interpolated height value. In our implementation, a Shepard interpolation method is applied:

$$d = \sum_{i=0}^{n-1} (1/r_i)^2 \cdot d_i \Big/ \sum_{i=0}^{n-1} (1/r_i)^2$$

where d is the height of an arbitrary point P within the geographical boundary, d_i are the known heights (attributes) at the known points C_i (*e.g.* center points of zip codes or counties), and r_i are the distances between P and C_i. A 2D image of the geographical map is used to limit the surface within the geographical border. This technique is implemented as a variation of our previous work on GeneTerrain [33].


(a)

(b)

Figure 5. Terrain views of a multi-disease visualization over the Indiana state map. (a) County based textures and interpolation; (b) County textures and zip-code based interpolation.

Figure 5 shows two examples of terrain surfaces mapped with noise pattern textures. Figure 6 shows a terrain surface example using offset contour texture for multi-diseases data. Figure 7 shows two terrain surface examples using offset contour textures for time-series data.

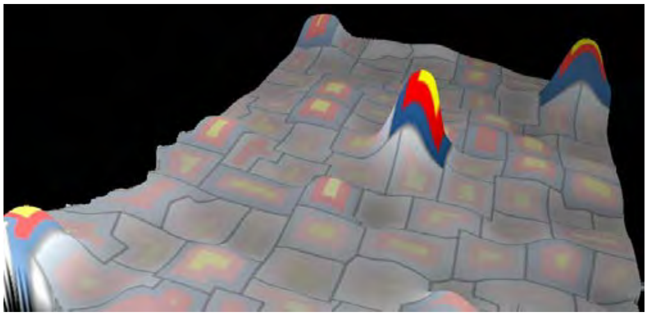
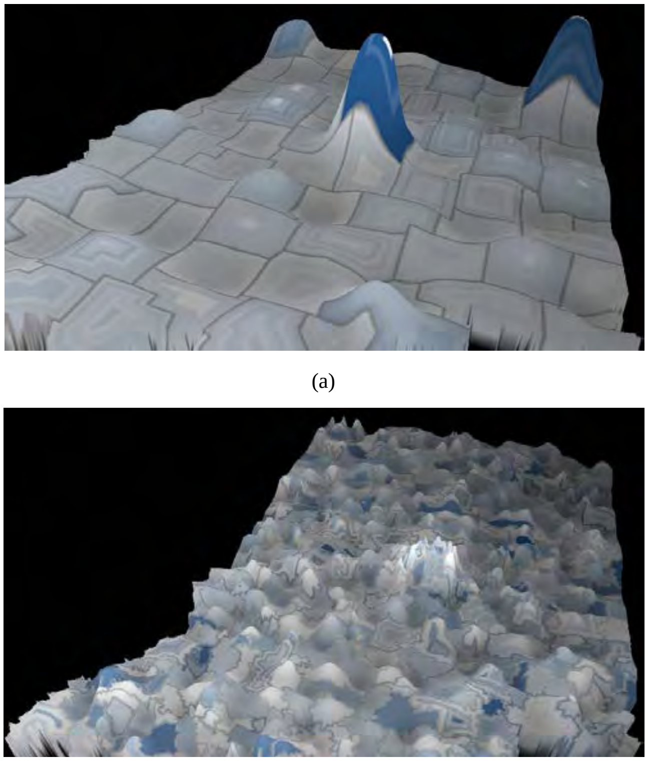


Figure 6. A terrain view of a multi-diseases visualization using a county-based offset contour texture


(a)

(b)

Figure 7. Terrain views of a time-series data over the Indiana state map. (a) County based; (b) zip-code based.

4.3 Ring Graph

Texturization and terrain surfaces provide overviews of health care data associated with geographic regions. In order to view more detailed patient level data, we also developed a new patient visualization method called Ring Graph. In Ring Graph, each patient is modeled as a point in a radial coordinate system. The radial space is subdivided into multiple rings, each of which represents one visualization term that was selected from the association map. These terms are typical disease names, but can also be other associated terms such as symptoms and risky behaviors. The circumference of this radial space represents the time-axis. Thus, time is encode as the radial angle of the points (patients). Ring Graph shows the distribution of patient-level data over a time-attribute space. One significant attribute, for example "age", will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots.

Occurrences of the same patient associated with multiple terms (e.g. diagnosed with multiple diseases) are connected with curves across the graph. A connecting curve will be highlighted when there is mouseover on the patient or the curve. Details of a patient records can also be shown by mouseover. To avoid clutter, the connecting curves are drawn with adjustable semi-transparent lines. Lowering the transparency can reveal more clearly the associations between terms. Figure 8 show an example of the Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis over a time period.

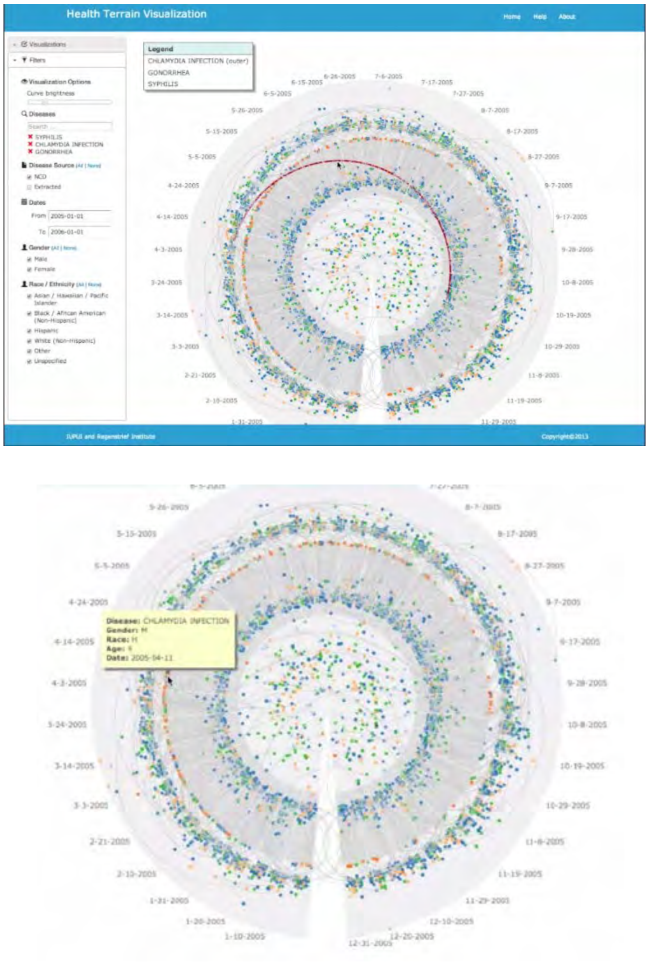


Figure 8: A Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis. For each patient (dot), the color represents race, the shape represents gender, and the radius represents age.

5 Conclusion

We presented a new geospatial visualization approach based on a texturization technique for large healthcare datasets. This approach takes advantages of the available geospatial space in healthcare visualization applications to visually represent multiple attributes or time-series data as texture patterns mapped onto the surfaces of geographic regions. This allows the users to visually filter and compare different color bands and data evolution patterns within a geographic context. An additional Ring Graph method can be applied as a supplementary technique for more detailed patient level visualization. These techniques have been integrated into our HelthTerrain visualization system for public health data visualization applications.

In the future, we would like to continue refining the visualization tools developed using these visualization techniques, as part of the HealthTerrain system. We would also like to develop a scalable terrain surface method to dynamically adjust the geospatial regions at different scales. It may also be beneficial to apply a smoothing filter to the color layers of the offset contours to generate smooth color transitions over time.

Acknowledgements

We gratefully acknowledge the contribution of Thanh Minh Nguyen, Anand Krishnan, Weizhi Li, Jennifer Williams; and the Regenstrief Institute's data engineers. The work is supported by the Department of the Army; award number W81CWH-13-1-0020.

References

[1] Grossman C, Powers B, McGinnis JM (Ed). Digital infrastructure for the learning health care system: the foundation for continuous improvement in health and health care. The National Academies Press, 2011

[2] Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B., Lifeline: Visualizing Personal Histories, CHI, 1996, pp. 221-227.

[3] Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records, CHI'08, 2008, pp. 457-466.

[4] Bui, A., Aberle, D.R., Kangarloo, H. Timeline: Visualizing Integrated Patient Records. IEEE Trans. Information Technology in Biomedicine 11(4):462-473.

[5] Mane, K., Borner, K. Computational Diagnostics: A Novel Approach to Viewing Medical Data. Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV '07, 2007, pp. 27-34.

[6] Hallett, C. Multi-Modal Presentation of Medical Histories. IUI'08: 13th International Conference on Intelligent User Interfaces. 2008, pp. 80-89.

[7] Carroll LN et al. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.04.006

[8] GEMMELL J., ARIS A., LUEDER R.: Telling stories with MyLifeBits. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on (2005), IEEE, pp. 1536–1539.

[9] THE NEW YORK TIMES COMPANY: Openpaths, Feb. 2013. URL: https://openpaths.cc.

[10] GOOGLE: Latitude, Feb. 2013. URL: http://www.google.com/latitude/.

[11] ECCLES R., KAPLER T., HARPER R., WRIGHT W.: Stories in GeoTime. In VAST (Oct. 2007), Ieee, pp. 19–26.

[12] M. Kraak, "The Space Time Cube Revisited from a Geovisualization Perspective," Proc. 21st Int'l Cartographic Conf., pp. 1988-1996, 2003.

[13] Kraak, Menno-Jan, and P. F. Madzudzo. "Space time visualization for epidemiological research." ICC 2007: Proceedings of the 23nd international cartographic conference ICC: Cartography for everyone and for you. 2007.

[14] Kraak, M. J. and A. Kousoulakou (2004). A visualization environment for the space-time-cube. Developments in spatial data handling 11th International Symposium on Spatial Data Handling. P. F. Fisher. Berlin, Springer Verlag: 189-200.

[15] Kwan, M. P. (2000). "Interactive geovisualization of activity travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set." Transportation Research C 8: 185-203

[16] Andrienko, N., G. L. Andrienko, et al. (2003). Visual data exploration using space-time cube. 21st International Cartographic Conference, Durban, South Africa.

[17] B. Cabral and L. C. Leedom. Imaging Vector Fields Using Line Integral Convolution. In Poceedings of ACM SIGGRAPH 1993, Annual Conference Series, pages 263–272, 1993

[18] D. Stalling and H. Hege. Fast and Resolution Independent Line Integral Convolution. In Proceedings of ACM SIGGRAPH 95, Annual Conference Series, pages 249–256. ACM SIGGRAPH, 1995.

[19] R. S. Laramee, H. Hauser, H. Doleisch, F. H. Post, B. Vrolijk, and D. Weiskopf. The State of the Art in Flow Visualization: Dense and Texture-Based Techniques. Computer Graphics Forum, 3(2):203–221, June 2004.

[20] T. McGraw, M.Nadar. Fast Texture-Based Tensor Field Visualization for DT-MRI. 4th IEEE International Symposium on Biomedical Imaging: Macro to Nano, pp. 760-763, 2007.

[21] Cornelia Auer, Claudia Stripf, Andrea Kratz, Ingrid Hotz. Glyph- and Texture-based Visualization of Segmented Tensor Fields. Proc. Int. Conf. on Information Visualization Theory and Applications, 2012, 670-677.

[22] Daniel A. Keim, "Designing Pixel-Oriented Visualization Techniques: Theory and Applications," IEEE Transactions on Visualization and Computer Graphics, vol. 6, no. 1, pp. 59-78, January-March, 2000

[23] D.A. Keim and H.-P. Kriegel, ªVisDB: Database Exploration Using Multidimensional Visualization,º IEEE Computer Graphics & Applications, pp. 40-49, Sept. 1994

[24] D.A. Keim, H.-P. Kriegel, and M. Ankerst, ªRecursive Pattern: A Technique for Visualizing Very Large Amounts of Data,º Proc. Visualization '95, pp. 279-286, 1995.

[25] M. Ankerst, D.A. Keim, and H.-P. Kriegel, ªCircle Segments: A Technique for Visually Exploring Large Multidimensional Data Set,º Proc. Visualization '96, 1996.

[26] T. Lammarsch, W. Aigner, A. Bertone, J. Gartner, E. Mayr, S. Miksch, and M. Smuc. Hierarchical Temporal Patterns and Interactive Aggregated Views for Pixel-based visualization. 13th International Conference on Information Visualization, 2009, 44-50.

[27] Overhage JM, Grannis SJ, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. Am J Public Health. 2008 Feb;98(2):344-50. PubMed PMID: 18172157.

[28] Stephen G. Kobourov. Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv: 1201.3011.

[29] Nathan Gossett, Baoquan Chen. Paint Inspired Color Mixing and Compositing for Visualization. IEEE Symposium on Information Visualization 2004. 113-117.

[30] Ken Perlin. An image synthesizer. In Proceedings of SIGGRAPH85, pages 287–296. ACM Press, 1985.

[31] Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.

[32] Hoschek, J., (1988), “Spline Approximation of Offset Curves,” Computer Aided Geometric Design, Vol. 5, pp. 33–40.

[33] You, Q., Fang, S., Chen, J. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks. Journal of Information Visualization, 2010; 9:1, 1-12.

Health-Terrain: A Visual Analytics System for Health Data

Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], Yuni Xia, PhD[1], Sam Bloomquist[1], Thanh Minh Nguyen[1], Anand Krishnan[2], Shenghui Jiang[1], Weizhi Li[2], Jeremy Keiper[3], Shaun Grannis, MD[3]

[1] Department of Computer Science, Indiana University – Purdue University, Indianapolis;

[2] School of Informatics and Computing, Indiana University – Purdue University, Indianapolis;

[3] Regenstrief Institute, Indianapolis, IN

Address correspondence to:

Shiaofen Fang, Ph.D.

723 W Michigan St. SL 280

Indianapolis, IN 46202

Phone: (317) 274-9731

Fax: (317) 274-9742

sfang@cs.iupui.edu

ABSTRACT:

OBJECTIVE: We aim to develop a visual analytics system by integrating information visualization, web based user interaction, text mining, and data mining techniques. To test the effectiveness of the visualization approach, this system will be applied on a use case based on a Notifiable Condition Detector (NCD) system developed by Regenstrief Institute.

METHODS: This visualization framework integrates concept space based knowledge extraction and information visualization. Raw health-related data is first processed and transformed to an information-rich health concept space, and then visualized using information visualization techniques. A new health-terrain surface technique, enhanced with attribute-encoded textures, is employed to generate a comprehensive visual representation for multi-dimensional and time-series data on geographical regions.

RESULTS: A browser-based prototype system has been developed for the interactive visualization of large health care data. It provides a content rich and interactive environment for real time decision making and trend/pattern detection. The system is being tested on an NCD dataset.

DISCUSSIONS: Data and text mining tools are applied to effectively convert data to a unified information platform – the concept space. This is the key to a general visualization framework for health data. A unique characteristics of public health data visualization is the need to embed geographical information in the visualization, which motivates our design of the health-terrain algorithms.

CONCLUSIONS: The visualization system offers a real time solution for the effective use of large scale electronic health record systems by allowing system level integration of the human´s visual capabilities into the overall health data based decision making system.

KEYWORDS: information visualization; visual analytics; public health data, notifiable condition detector, text mining; data mining.

BACKGROUND AND SIGNIFICANCE

Personal health records, point of care testing systems, and other ancillary digital devices have the potential to generate massive amounts of digital health related information. Consequently, clinical providers and health care administrators must discern meaningful patterns from increasingly large datasets to make informed clinical decisions for individual patients and develop broad strategies for patient populations [1]. Variations in data types and the diversity of data consumers increase the complexity of the challenge. To effectively leverage health information to support timely decision-making and enable effective trend/pattern detection, increasingly innovative and novel methods for visualizing healthcare data using pragmatic and easily interpretable frameworks are necessary. Several challenges must be addressed to achieve this goal:

(1) Complexity. Health data is a data-rich, information-poor domain. In Electronic Health Record (EHR) systems, data are heterogeneous, unstructured, hierarchical, and longitudinal.

(2) Large Scale. EHR systems are often extremely large. While it is possible to visualize an EHR system in a limited scope, high impact knowledge discoveries more likely come from global scale (population-wide) visualization and knowledge mining.

(3) Spatiotemporal needs. Visualizing population-level health data often involves presenting geospatial and time-series data in a common visual context. This presents a challenge in visual encoding of the information space.

Health data visualization of large-scale datasets has not been extensively studied. There are several existing works and visualization systems that deal with the secondary use of electronic health record data in a limited scope. *LifeLines* [2] uses a traditional 2D time line visualization technique to visualize specific patient medical and health history. It emphasizes the visualization of temporal ordering of events with limited aggregation effect. An extension of *LifeLine, LifeLine2* [3], enables multiple patient comparisons and aggregation for analysis, but the visualization design limited its scalability. A similar system, call *TimeLine* [4], re-organizes and re-groups multiple EHR content types in a layout of Y-axis to track multiple events along the same time line. A set of visualization tools are described in [5] for visualizing a patient's electronic health record to aid physicians diagnosis and decision-making. The traditional matrix view and parallel coordinates are the main techniques applied. *CLEF* [6] is a system enabling visual navigation through a patient's medical record using semantically and temporally organized networks to represent events throughout the patient's medical history. *CLEF* also supports limited text processing capabilities for generating textual summaries. None of these existing systems is capable of visualizing large-scale integrated EHR datasets. A review paper on visualization tools for infectious diseases is given in [7]. Visualization tools for syndromic surveillance activities were discussed in [8,9], and the associated evaluation and validation methods were discussed in [10,11].

There have also been many techniques for geospatial visualization of time-series data, such as color-coding of time [12], connecting time-lines [13], and time-curves [14], which often introduce visual clutter and occlusion. A well-known technique is Space-Time-Cube [15,16,17]. While time and spatial information are integrated in a 3D visual representation in a space-time-cube, the sense of space-time embedding diminishes as the data moves up in the time axis since the geospatial map is only given at time zero. Visual clutter will also be a problem with large datasets.

To address these shortcomings, we developed the HealthTerrain system. To test our HealthTerrain visualization system we used a large public health notifable disease reporting system. The Regenstrief Institute implemented and maintains an unparalleled HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in the State of Indiana. The Notifiable Condition Detector (NCD) System uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC) [18]. The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations [19]. The system automatically detects positive cases of notifiable conditions and forwards alerts to local and state health departments for review and follow up. These alerts enable more effective and efficient public health population health monitoring and case management.

OBJECTIVE

Our goal is to develop a prototype system supporting visualization and visual analytics of large healthcare data sets. The system integrates information visualization, web-based user interaction, text mining, and data mining techniques. It provides data exploration functionalities through interactive 2D and 3D visualization tools. The Notifiable Condition Detector (NCD) system is used to test the effectiveness of the system.

METHODS

The health-terrain visualization system demonstrates the feasibility of implementing two novel strategies:

(1) Implementing a common “Concept space” enables an integrated visualization platform. Raw health data are first processed by text and data mining algorithms, and converted to an information-rich concept space where new terms and their relationships are represented in an association map, as a space of extracted partial knowledge. The concept space uses a controlled vocabulary that can be pre-defined based on application needs, and enhanced by data/text mining algorithms. Therefore, our evaluation suggests that this framework is scalable and can be expanded progressively as the applications and use cases expand.

(2) Spatiotemporal visualization. Population health data visualizations commonly seek to represent geospatial information and time-series information within the same visual context. We developed a spatiotemporal visualization technique to accommodate multi-dimensional and time-series health data using 3D terrain surfaces. Surface texture patterns over a geographical map are color-coded using an offset contour technique to represent multiple attributes and changes over the time axis.

Concept Space Definition

The “concept space” represents a uniform layer of clinical observations and their associations, and enables users to explore data using various visualization and analysis methods. Concept terms are derived from data mining and text-mining processes applied to the use case datasets. The NCD dataset contains 833,710 public health notifiable cases spanning more than 10 years from among 439,547 unique patients. An additional dataset containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories were extracted from the Indiana Network for Patient Care (INPC) health information exchange for these patients. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including disease, symptom, mental behavior, and risky behavior. An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology.

Database Design

A 3-layer database model is designed to store the processed NCD dataset and the associated text data to support the queries for various visualization approaches. The first layer contains 4 base tables for patient, disease, term and location. The term table has 4 subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains the associations between different entities, and the third layer contains indirect associations between disease, location, and the other terms, which are constructed using data mining techniques.

Text Mining Process

We processed 325,791 unstructured clinical notes containing patient discharge summaries, laboratory reports, and medical histories using NLP techniques. We transformed the clinical notes from XML format to simple text format and also applied sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information. This was done with the help of the Unified Medical Language System (UMLS) [20], database repository of clinical and health related terms. After extracting the entities using NER, negation analysis was applied using NEGEX algorithm[4] to remove negated terms. The extracted terms contained lexical variants of the same diseases, symptoms, etc. To normalize these variations and de-duplicate terms we applied stemming and concept clustering algorithms [21].

We performed additional analysis to compute the correlations among conditions using the *tf-idf* (term frequency – inverse document frequency) vector space model to identify the significantly co-occurring diseases [22]. An association weight/score is also calculated for each pair of terms [23].

Data Mining Algorithms

Association Mining. We next compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities *i* and *j*, the rule confidence between *i* and *j* is computed as

$$\text{Rule\_confidence}(i, j) = \frac{\left|i \wedge j\right|}{|\, i \,|} \; 9$$

in which $\left|i \wedge j\right|$ is the number of patients showing both entities *i* and *j* and | *i* | is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time.

Clustering Analysis. We develop a co-clustering algorithm to cluster both diseases and other terms from text-mining to discover potential combinations of both diseases and other terms, potentially representing disease subtypes or new biomedical patterns. The algorithm iteratively and partially allocates the diseases or terms into clusters based on the rule confidence attributes.

Visualization Algorithms

Association Map. Association map (Figure 2) is a graph visualization of the association relationships among the diseases and other terms in the concept space. It can serve as a platform supporting interactive selection of concepts to dynamically visualize data using a variety of tools in the visualization system. To draw an association graph, a spring-embedder algorithm [24] is used to layout the graph nodes:

$$E_s = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}k(d(i,j) - s(i,j))^2$$

where *d(i,j)* is the 2D Euclidean distance of two nodes, and *s(i,j)* is a similarity metric of two nodes representing the heuristic of the layout. Edge thickness indicates the strength of association, and node size can reflect the number of other nodes to which a given node has a significant association, or the total occurrence of a term (e.g. disease) in the dataset. The graph highlights related nodes and the association edges that have significant associations to the selected index node.

Theme River. Theme river view [25] (Figure 3) shows the aggregate trend for the terms (e.g. diseases and symptoms) selected by the user for a given time period. Each term (a theme) is visually represented as a river stream, and implemented as a filled curve plot along the horizontal time axis, with y-axis representing the occurrence of the term. Multiple themes are stacked together vertically for side-by-side comparison of the streams over time, as well as the possible interactions.

Patient Graph. The Patient Graph (Figure 4) is a new visualization method designed specifically for patient-centric health data visualization. It shows the patient information distributed over a radial space highlighting time and disease dimensions. Occurrences of the same patient diagnosed with multiple diseases are connected with arcs across the graph. Alternating bands of gray and white represent different diseases, and time increases as one moves along the arc of the concentric circles. Within each disease band, the radius of the circle represents the age of the patient. Other variables such as race and gender are indicated by the color and shape of the nodes. Patient details are shown interactively with Mouseover windows.

Terrain Surface. A heat map view can be applied to a surface map to show the occurrence frequencies and distribution of a given term over a geographical area (e.g. the distribution over counties or zip-codes in the state of Indiana). Queries can be performed interactively over the sub-regions of the map. A 3D terrain surface can then be constructed by converting the heat map to a height field. This would allow us to color code an additional attribute on the 3D surface. Since only a small set of points (e.g. zip code centroids) have height values on the map, interpolation is necessary to construct a smooth surface. We applied a Shepard interpolation method:

$$d = \sum_{i=0}^{n-1} (1/r_i)^2 \cdot d_i \Bigg/ \sum_{i=0}^{n-1} (1/r_i)^2$$

where d is the height of an arbitrary point P within the Indiana map, di are the known heights (attributes) at the known points Ci (e.g. center points of zip codes or counties), and r_i are the distances between P and Ci. A 2D image of the state map is used to limit the surface to be within the Indiana border. Once the surface is constructed, horizontal cross sectional contours can be generated to identify the geographical regions that response more sensitively to the given term. This technique is implemented as a variation of our previous work on GeneTerrain [26].

Offset Contours on Terrain Surface. In order to visualize multiple attributes for a disease (e.g. the associated diseases or symptoms) in each sub-region on a terrain surface, a new offset contour technique is developed (Figure 5). A geographical region is subdivided into multiple sub-regions of similar shapes by offsetting the boundary curve toward the interior of the region, creating multiple offset contours. There are several geometric techniques to generate offset curves. Since we use the SVG W3C imaging operations to implement geographical maps, we applied SVG’s build-in image erosion operator [27] to generate the offset contours. Using different color patterns or textures in these different sub-regions is an effective way to represent multiple attributes for sub-regions.

Using the same technique, we can visualize time-series data on geographical regions (Figure 6). The time line can be divided into multiple time intervals which are represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time.

This approach is particularly suitable for population level healthcare data, which typically have attributes defined for geographical regions (counties or zip-codes). Although the precise distribution of the cases are not known, the smooth interpolation provides some level of reasonable spread across the geographical regions.

System Integration and Interaction

We designed and implemented the framework of the HealthTerrain visualization system using WebGL in an HTML5 canvas. The system’s architecture pattern is based on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture.

The user interface is a modern web GUI using a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats such as Extensible Markup Language (XML) and JavaScript Object Notation (JSON). The visualizations use HTML, CSS, SVG, and WebGL technologies with a number of open-source Javascript libraries such as sigma.js, d3.js, jquery.js and three.js for drawing, displaying and interacting with data and graphics. The interactive user interface is designed to support data exploration and hypothesis generation. The user interaction is based on concept associations. The process works as follows:

1) The user picks the concept terms that are of interest to the user's visualization objectives from the association map (Figure 2). A data filter will then be selected using the filter interface (Figure 4), which includes time, gender, race, age, etc.
2) Based on the type of combination of the concept terms, the appropriate visualization tools will be activated to visualize the filtered dataset. Depending on the type of visualization tools, interactive operations, such as zooming, mouseover, picking, and 3D rotation, can be applied by the user to achieve various visualization effects. In particular, multiscale visualization can be applied when zooming in to a geospatial region by switching the visualization from a county-based visual representation to a zip-based visual representation.
3) After exploring a visualization method, the user may want to generate another related visualization for comparison or a follow-up data examination. The current visualization in the display window can be minimized to the sidebar column of the primary system window (Figure 2a), which can later be activated again as needed.
4) We are currently implementing a new function that allows the user to group multiple concept terms into a composite term. This effectively create a visual analytics loop for generating new concept and patterns.

RESULTS

Data and Text Mining

Data mining and text mining techniques are first applied to generate the concept space, and visualization algorithms are then applied using the concept space to generate the visualization results and user interactions.

After initial processing, we apply NER to the unstructured reports using UMLS to identify diseases, symptoms, mental behavior, risky behavior and other medication terms from the 325,791 reports. The total number of terms extracted for each category is given in Table 1. The *tf-idf* model is used to identify comorbidity of the diseases across the 325,791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications). Table 2 shows the top 10 diseases and the corresponding conditions.

Table 1: Total terms identified by NLP

Term Type	**Number of terms extracted using NLP**
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

Table 2**:** Comorbid conditions with top 10 most occurring diseases

	Diseases	**Symptoms**	**Mental Behavior**	**Risky Behavior**	**Medications**
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, smokes, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, type 2, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, compliant, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion, altered mental status,	smoking, impression, drinking, smokes, tobacco, compliant, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, smokes, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	prn pain, constipation, cramping, headache, breakthrough pain,	abuse, drug use, depression, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused	smoking, compliance, impression, tobacco use, illicit drug use, drinking	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infection, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

The database contains 439,547 patients, 1,976 diseases, 3,756 locations and 3,851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2,847 medications). At the second layer, the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. A total of 90,376 patients were associated with at least one non-disease term. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term. Figure 1 shows the most frequent diseases in the NCD dataset.

Figure 1. Most Common Diseases in the NCD dataset

Visualization

The Health-Terrain visual analytics system is a browser based interactive system that can be used as an exploratory tool for public health administrators and clinicians to visually explore and analyze patient-based healthcare datasets. Figure 2 shows the system interface with the disease association map. A visualization usually starts with an association map where the users select the diseases or other terms to be visualized using the various visualization tools. Figure 3 show a Theme River view of total occurrences of Hepatitis A, B, C and D over time. Since the incidence of Hepatitis D is very small, its theme flow is too thin to see in the graph. Figure 4 shows the patient graph for the same 4 diseases. The innermost circular region represents Hepatitis D, which has very few cases. The next circular band represents Hepatitis A which has a fairly even age distribution except that there is a concentration of young black patients. Hepatitis C (the next band) shows a heavy concentration of middle-age patients. A cross-arc represents the same patient diagnosed with multiple diseases at different times. Figure 5 shows an example of the offset contour visualization of multiple diseases over the Indiana mapfor the years of 2008 and 2009. The cases are highly concentrated in Central Indiana (Marion and surrounding counties). A time-series example is shown in Figure 6. It includes all the cases of Lyme disease from January 2008 to December 2009. The time period is divided into 6 subintervals. The shades of blue color are used to represent different levels of the occurrence in each county. Again, due to the population imbalance, the cases are heavily focused in Central Indiana.

Figure 2. (a) An association map of all diseases; (b) a close-up view with user selections.

Figure 3. (a) A heat map view over the Indiana state map; (b) Theme River view of Hepatitis A, B, C and D. The casese for Hepatitis B is too small to be seen in the graph.

Figure 4. Patient graph for Hepatitis A, B, C and D. For each patient (dot), the color represents race, the shape represents gender, and the radius represents age. Cross-arcs represent the occurrences of multiple diseases diagnosed for the same patients.

Figure 5. Multi-attribute Terrain Surfaces over Indiana map. (a) County based occurrences; (b) Zip-code based occurrences.

Figure 6. Time-series Terrain Surfaces over Indiana map. (a) County based occurrences; (b) Zip-code based occurrences.

DISCUSSIONS

The Health-Terrain visualization system provides an interactive browser-based platform for data exploration and visual analytics of large healthcare data, and contributes meaningfully to the body of technical knowledge related to visualizing large-scale clinical data sets, which will in turn inform best practices related to initiatives seeking to leverage large-scale clinical data sets. This prototype system augments users' ability to discern patterns in large-scale data, ultimately leading to better-informed decisions by clinicians and managers for individual patients and populations.

The information-rich concept space effectively compresses large, heterogeneous, and historical patient and public health data into a unified, intuitive and comprehensive data representation. Any patient-based dataset can be easily converted into a concept space representation using a set of standard text- and data-mining tools, which can then be visualized by the system. Our future work seeks to provide more customizable interface features so that the system can be adapted to different healthcare applications. This allows the visualization system to operate independent of specific data formats, and can also help facilitate interoperability among multiple EHR systems.

The visualization techniques developed in this system are specifically designed to suit the interactive exploration of healthcare data. The Association Map provides an initial platform to explore relationships of various health concepts within a controlled ontology. The Patient Graph provides a means to explore temporal patterns within patient populations over time or other features. The terrain-based visualization technique is a promising approach for large health data as it is effective in revealing trends, patterns, and abnormalities. The offset contour texture approach is an innovative visualization technique for spatiotemporal data visualization, and provides a way for the users to visually filter different color bands within the context of a spatial landscape. This is particularly important for healthcare data which often involves the integration of geographical information and population level diseases information. The use of offset contours for time-series data is a unique and promising concept as it provides a non-animation based spatiotemporal visualization with rich geographical context.

With increasingly large collections of clinical and notifiable disease data we can explore many potential correlations between particular diseases and other clinical features (such as clinical concepts ground in discharge summaries); such correlations may reveal predictors of clinical outcomes and suggest potential future interventions to reduce disease burden. For example in previous a studies we identified specific communicable diseases that were associated with other rates of co-morbid communicable disease [27]. Using the previously-described Patient Graph, association map and Theme River visualizations we are able to explore potential correlations among communicable diseases (e.g., HIV and Syphilis), clinical concepts related to communicable disease (e.g., alcohol use and sexually transmitted disease), and temporal correlations among diseases. Preliminary Theme River visualizations suggest that particular patterns of community acquired infections may increase one’s risk of nosocomial infections.

To ensure usability for clinical and public health users, the current visualization framework has been designed with input from public health and clinical stakeholders. As part of an initiative currently funded by the DoD, we will leverage usability guidance [28,29] to assess workflow, alerts, navigation, and layout, and visualization effectiveness among additional public health and clinical stakeholders. These assessments will inform future revision of the framework.

CONCLUSIONS

A browser-based visualization system offers a real time solution for the effective use of large scale electronic health record systems by allowing system level integration of the human´s visual capabilities into the overall health data based decision making system. The visual representation of concept space provides a method to compress large, heterogeneous, and historical patient and public health data into a single, intuitive and comprehensive visualization, which can also facilitates interoperability among multiple electronic health record systems. The system described in this paper is in prototype form and we are planning to further system and software development to disseminate a general tool for healthcare administrators and clinicians.

ACKNOWLEDGEMENTS

The project is supported by the Department of the Army; award number W81CWH-13-1-0020. We gratefully acknowledge the contribution of Jennifer Williams; and the Regenstrief Institute's data engineers.

REFERENCES

1. Grossman C, Powers B, McGinnis JM (Ed). Digital infrastructure for the learning health care system: the foundation for continuous improvement in health and health care. The National Academies Press, 2011
2. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B., Lifeline: Visualizing Personal Histories, CHI, 1996, pp. 221-227.
3. Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B. Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records, CHI'08, 2008, pp. 457-466.
4. Bui, A., Aberle, D.R., Kangarloo, H. Timeline: Visualizing Integrated Patient Records. IEEE Trans. Information Technology in Biomedicine 11(4):462-473.
5. Mane, K., Borner, K. Computational Diagnostics: A Novel Approach to Viewing Medical Data. Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, CMV '07, 2007, pp. 27-34.

6. Hallett, C. Multi-Modal Presentation of Medical Histories. IUI'08: 13th International Conference on Intelligent User Interfaces. 2008, pp. 80-89.

7. Carroll LN et al. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.04.006.

8. Grannis SJ, Egg J, Overhage JM. Reviewing and managing syndromic surveillance SaTScan datasets using an open source data visualization tool. AMIA Annu Symp Proc. 2005:967. PubMed PMID: 16779254.

9. Maciejewski R, Rudolph S, Hafen R, Abusalah A, Yakout M, Ouzzani M, Cleveland WS, Grannis SJ, Wade M, Ebert DS. Understanding Syndromic Hotspots - A Visual Analytics Approach. IEEE Symposium on Visual Analytics Science and Technology, pp. 35-42, 2008.

10. Duke JD, Li X, Grannis SJ. Data visualization speeds review of potential adverse drug events in patients on multiple medications. J Biomed Inform. 2010 Apr;43(2):326-331. PubMed PMID: 19995616.

11. Maciejewski R, Hafen R, Rudolph S, Tebbetts G, Cleveland WS, Ebert DS, Grannis SJ. Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. IEEE Comput Graph Appl. 2009 May-Jun;29(3):18-28. PubMed PMID: 19642612.

12. THE NEW YORK TIMES COMPANY: Openpaths, Feb. 2013. URL: https://openpaths.cc.

13. GOOGLE: Latitude, Feb. 2013. URL: http://www.google.com/latitude/.

14. ECCLES R., KAPLER T., HARPER R., WRIGHT W.: Stories in GeoTime. In VAST (Oct. 2007), Ieee, pp. 19–26.

15. Kraak, Menno-Jan, and P. F. Madzudzo. "Space time visualization for epidemiological research." ICC 2007: Proceedings of the 23nd international cartographic conference ICC: Cartography for everyone and for you. 2007.

16. Kraak, M. J. and A. Kousoulakou (2004). A visualization environment for the space-time-cube. Developments in spatial data handling 11th International Symposium on Spatial Data Handling. P. F. Fisher. Berlin, Springer Verlag: 189-200.

17. Andrienko, N., G. L. Andrienko, et al. (2003). Visual data exploration using space-time cube. 21st International Cartographic Conference, Durban, South Africa.

18. Overhage JM, Grannis SJ, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. Am J Public Health. 2008 Feb;98(2):344-50. PubMed PMID: 18172157.

19. Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html

20. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11

21. S. Osinski, D Weiss. A concept-driven algorithm for clustering search results - Intelligent Systems, IEEE Intelligent Systems, May/June, 2005. pp. 48-54.

22. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/

23. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342

24. Stephen G. Kobourov. Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv: 1201.3011.

25. S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *Visualization and Computer Graphics, IEEE Transactions*, vol. 8, pp. 9-20, 2002

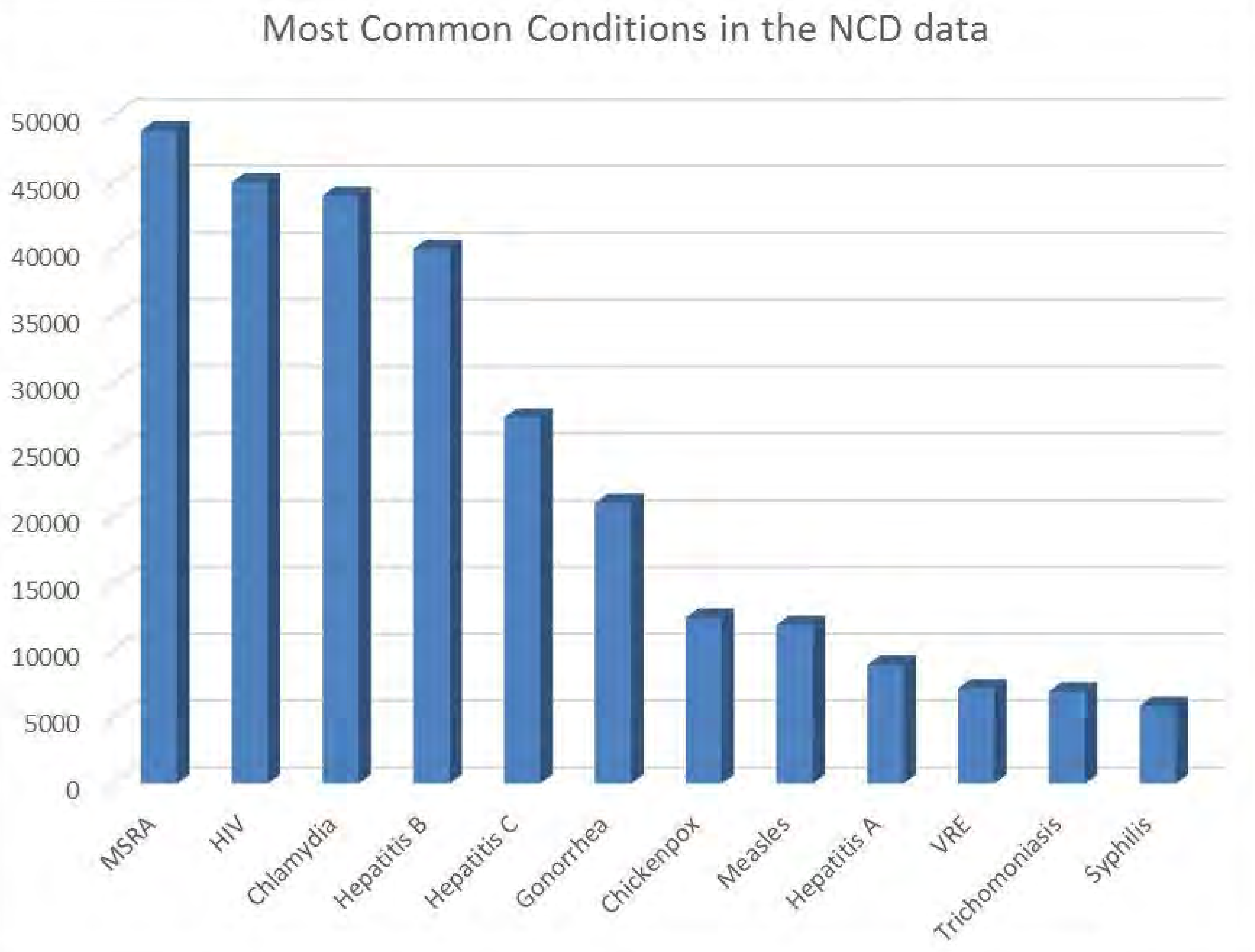
26. You, Q., Fang, S., Chen, J. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks. Journal of Information Visualization, 2010; 9:1, 1-12.

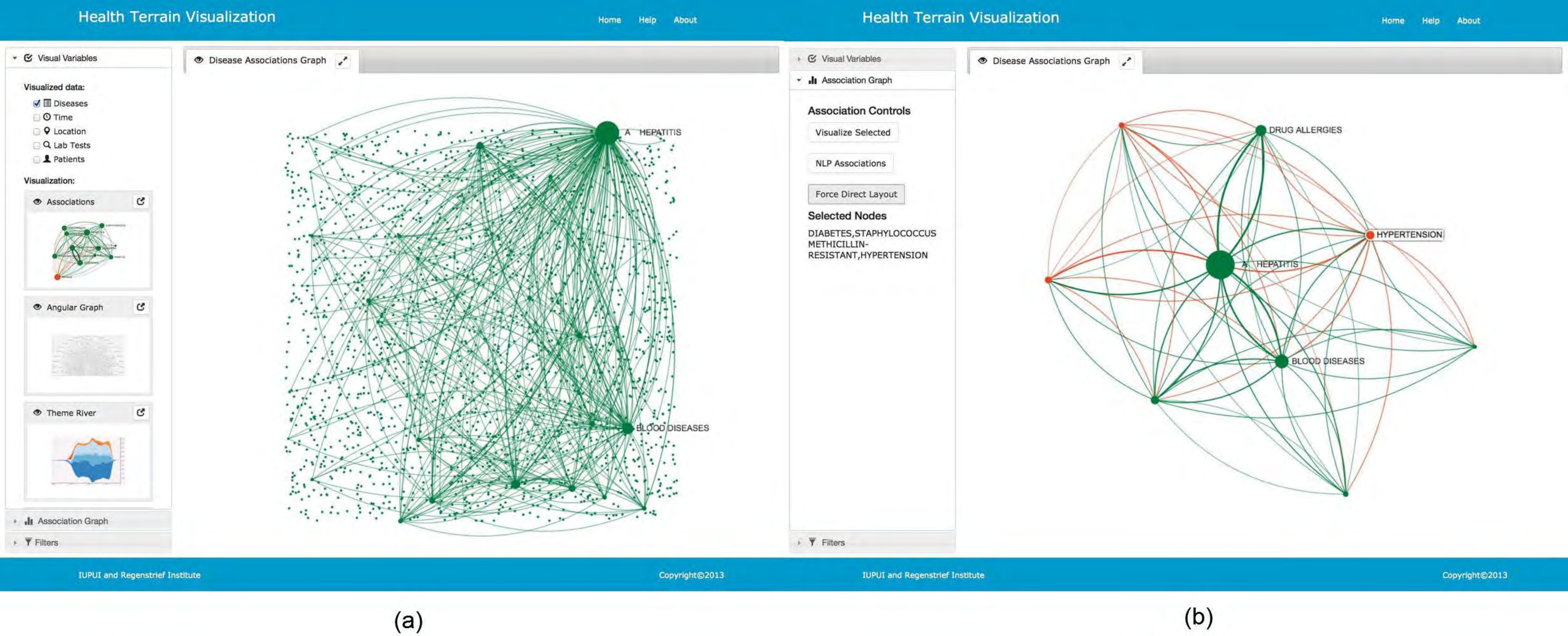
    Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.

27. Gichoya J, Gamache RE, Vreeman DJ, Dixon BE, Finnell JT, Grannis S. An evaluation of the rates of repeat notifiable disease reporting and patient crossover using a health information exchange-based automated electronic laboratory reporting system. AMIA Annu Symp Proc. 2012;2012:1229-36.

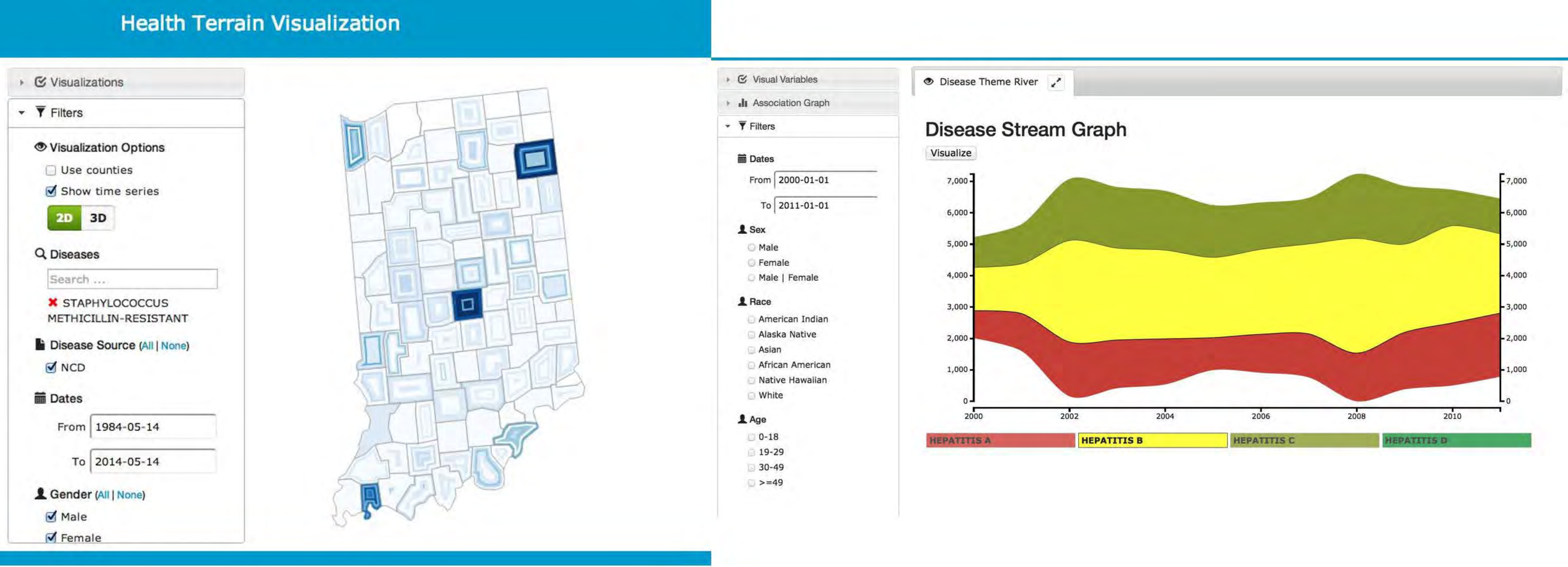
28. NIST Interagency/Internal Report - 7432. (2010) Common Industry Specification for Usability — Requirements. Retrieved July 10, 2014 from http://www.nist.gov/manuscript-publication-search.cfm?pub_id=51179.

29. Zhang J, Walji M. TURF: Toward a unified framework of EHR usability. Journal of Biomedical Informatics, 2011; 44 (6):1056-1067.


Most Common Conditions in the NCD data
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
MSRA
HIV
Chlamydia
Hepatitis B
Hepatitis C
Gonorrhea
Chickenpox
Measles
Hepatitis A
VRE
Trichomoniasis
Syphilis



DIABETES,STAPHYLOCOCCUS METHICILLIN-RESISTANT,HYPERTENSION



Health Terrain Visualization

- Visualizations
- Filters

Visualization Options

☐ Use counties

☑ Show time series

2D 3D

Diseases

Search ...

✖ STAPHYLOCOCCUS METHICILLIN-RESISTANT

Disease Source (All | None)

☑ NCD

Dates

From 1984-05-14

To 2014-05-14

Gender (All | None)

☑ Male

☑ Female

(a)

- Visual Variables
- Association Graph
- Filters

Dates

From 2000-01-01

To 2011-01-01

Sex

○ Male

○ Female

○ Male | Female

Race

☐ American Indian

☐ Alaska Native

☐ Asian

☐ African American

☐ Native Hawaiian

☐ White

Age

☐ 0-18

☐ 19-29

☐ 30-49

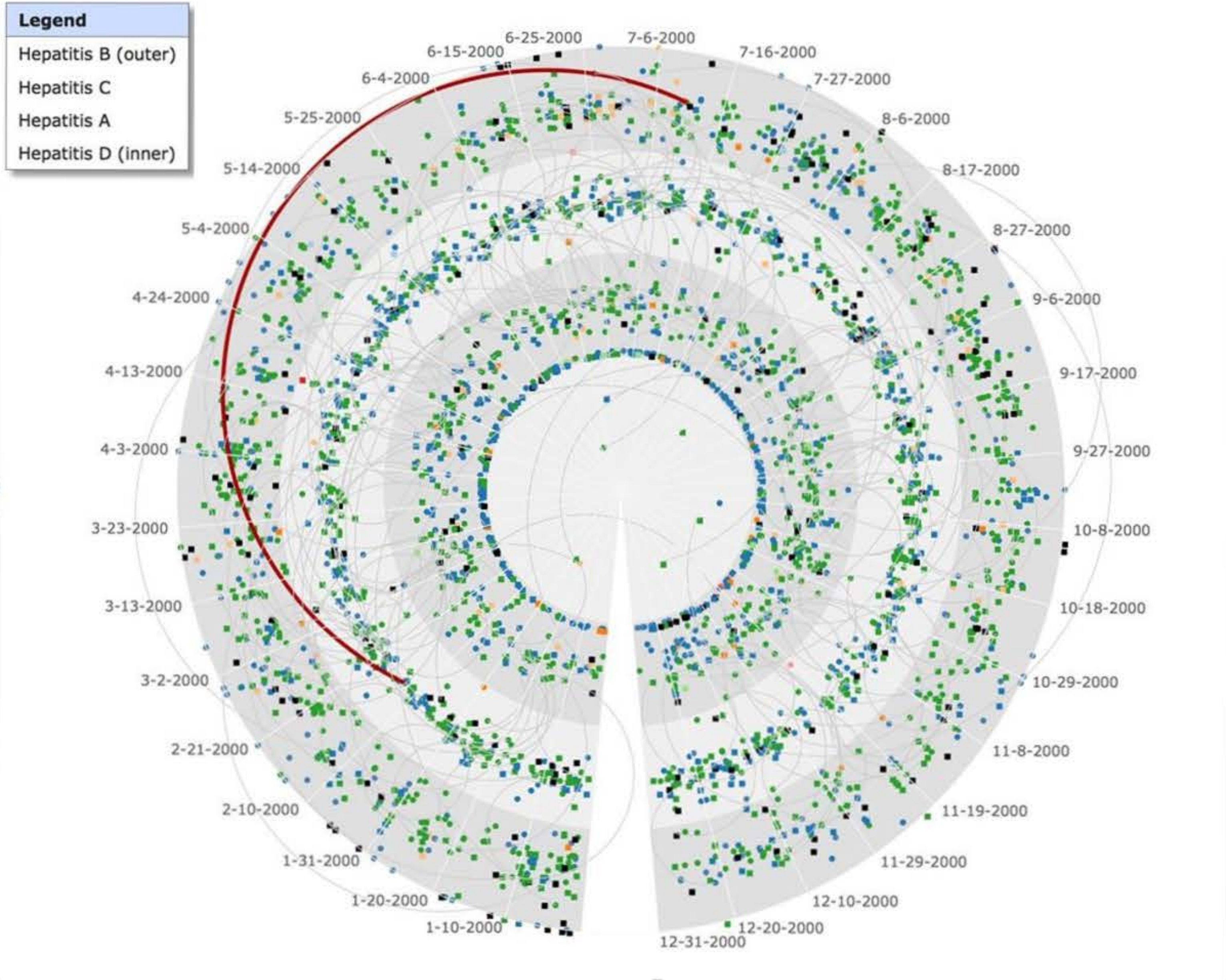
☐ >=49

Disease Theme River

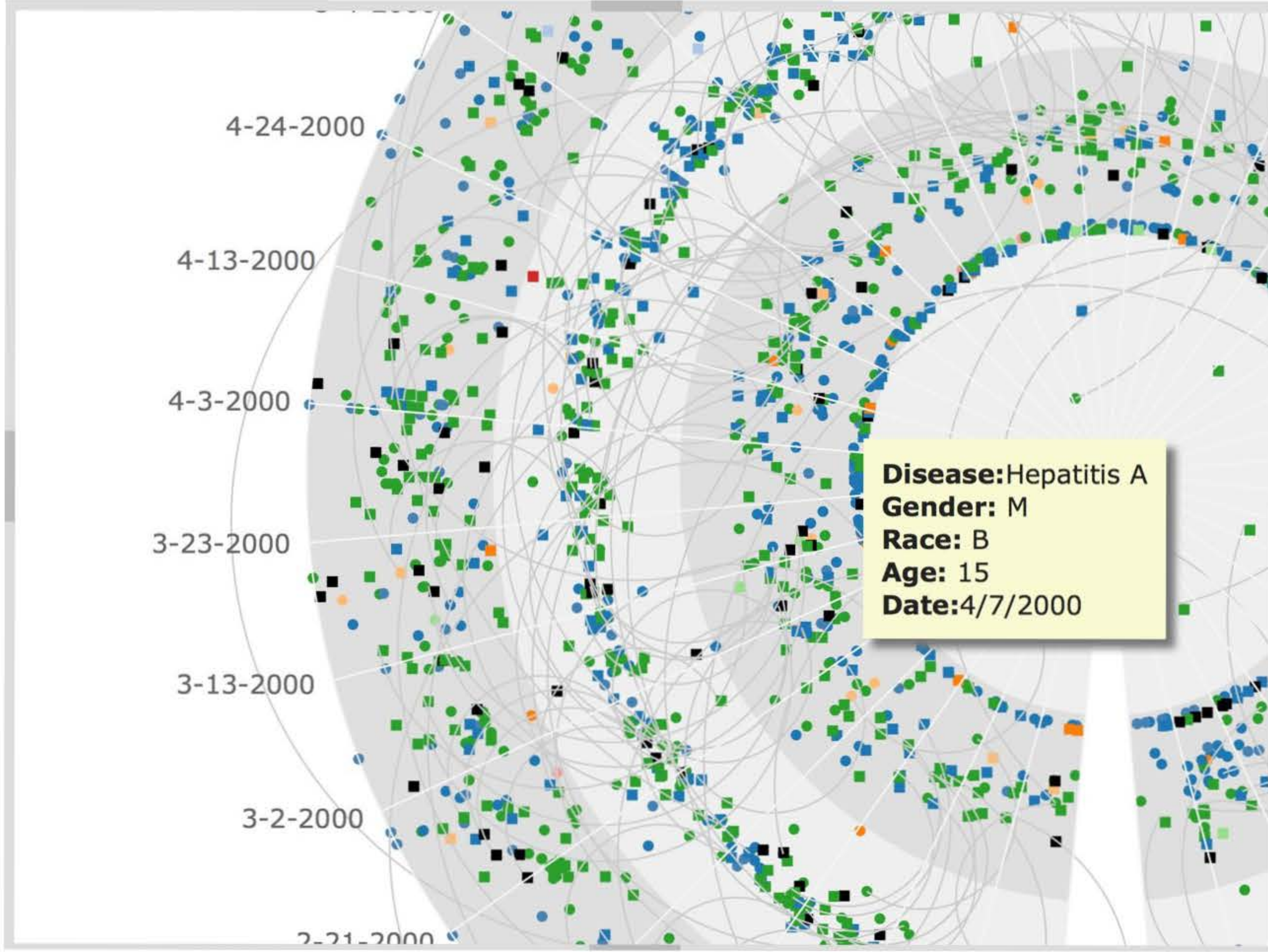
Disease Stream Graph

Visualize

(b)

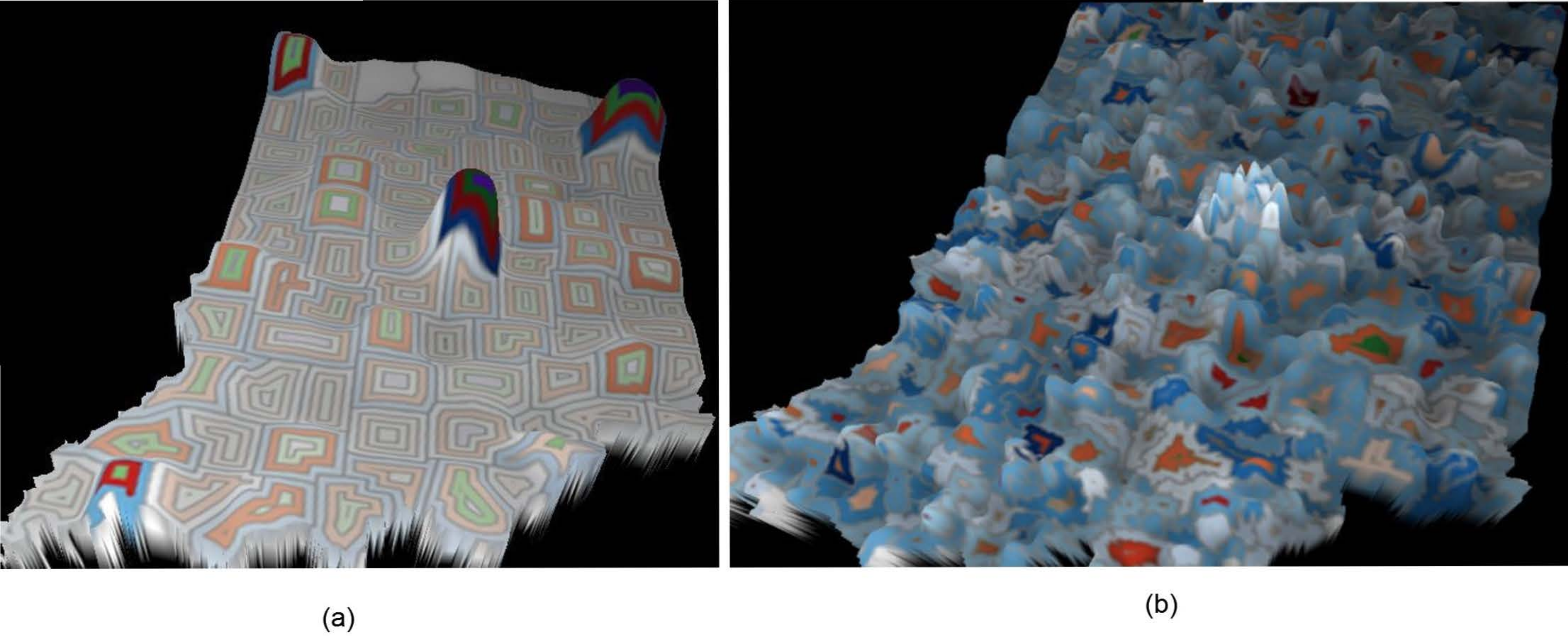


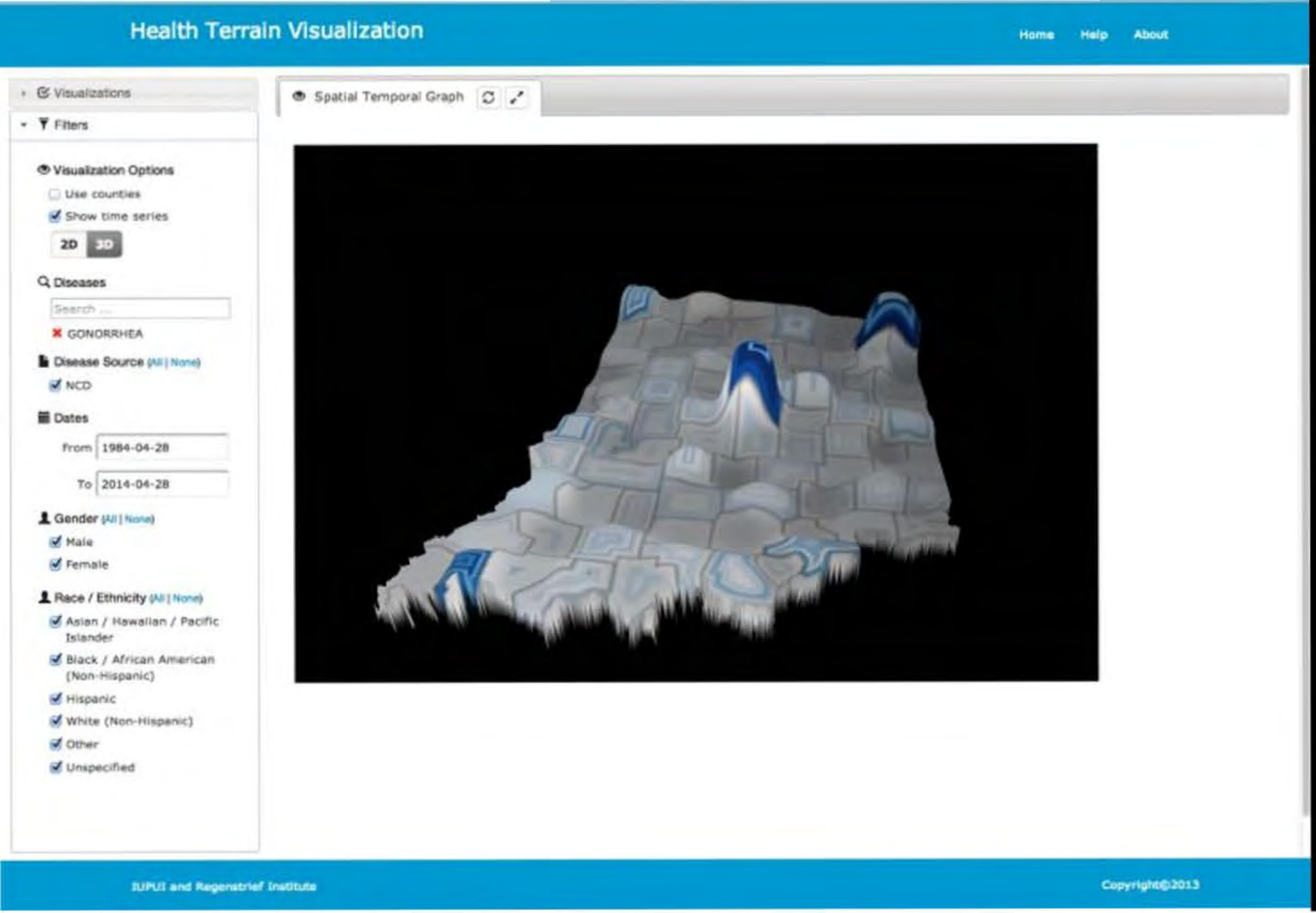
(a)



Disease: Hepatitis A
Gender: M
Race: B
Age: 15
Date:4/7/2000

(b)


(a)
(b)



Visualization Options

□ Use counties

☑ Show time series

Diseases

Search ...

✖ GONORRHEA

Disease Source (All | None)

☑ NCD

Dates

From 1984-04-28

To 2014-04-28

Gender (All | None)

☑ Male

☑ Female

Race / Ethnicity (All | None)

☑ Asian / Hawaiian / Pacific Islander

☑ Black / African American (Non-Hispanic)

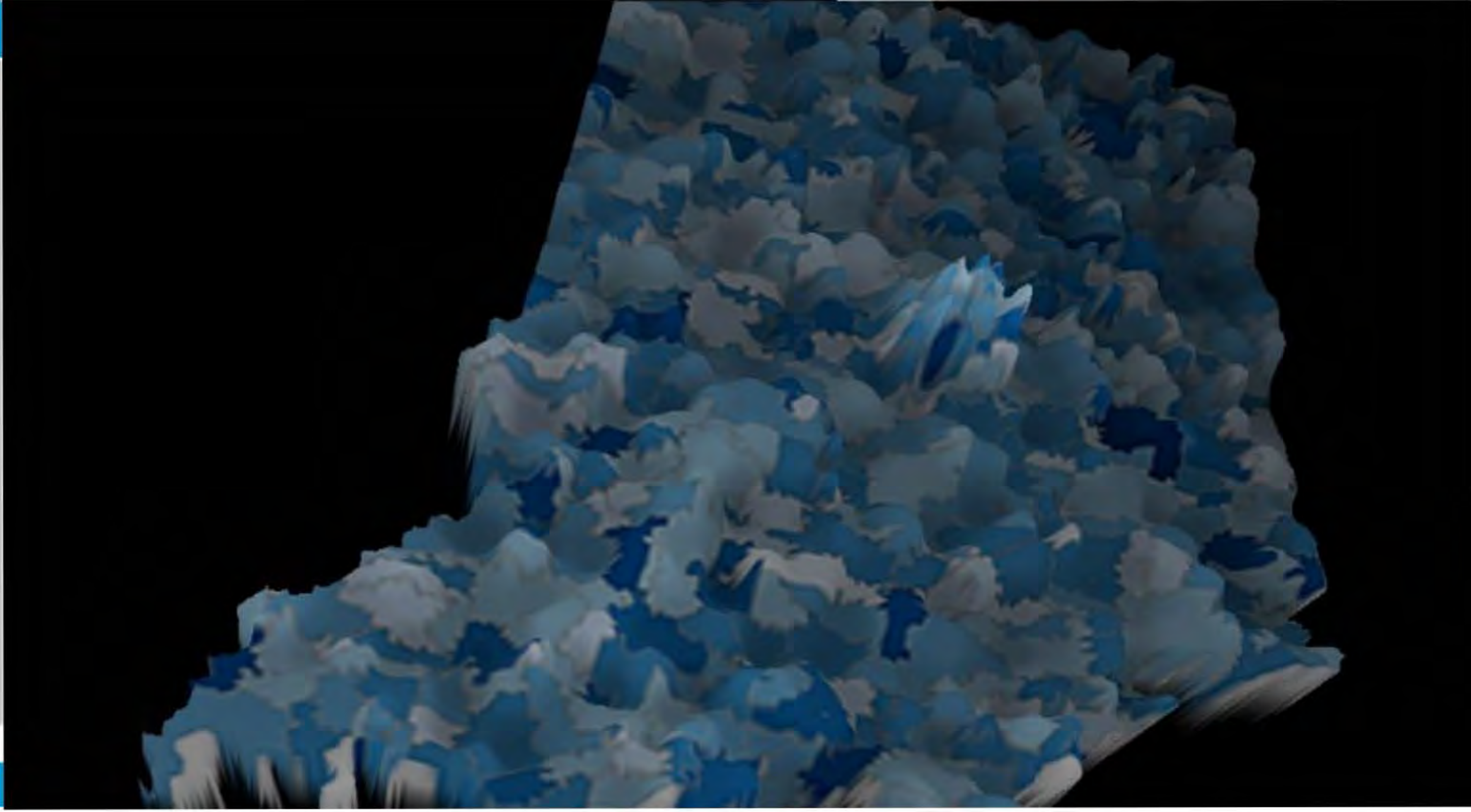
☑ Hispanic

☑ White (Non-Hispanic)

☑ Other

☑ Unspecified

(a)



(b)

Demonstrating A Public Health Terrain Data Visualization System

Jeremy Keiper[1], Shiaofen Fang, PhD[1], Yuni Xia, PhD[1], Mathew Palakal, PhD[2], Shaun Grannis, MD[3], Thanh Minh Nguyen[1], Sam Bloomquist[1], Anand Krishnan[2], Weizhi Li[1]

[1] Department of Computer Science, Indiana University – Purdue University, Indianapolis;

[2] School of Informatics, Indiana University – Purdue University, Indianapolis;

[3] Regenstrief Institute, Indianapolis, IN

Abstract

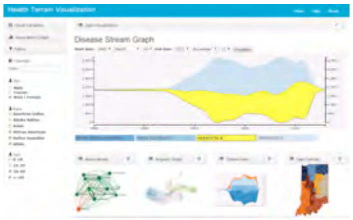
Use cases for public health data visualization systems indicate a need for an interactive system, a wide variety of potential filters for narrowing scope, and visualizations in both geospatial and logical domains. A public health study not only involves relationships between individuals within a given community, but also considers correlations among attributes of the diseases and mental behaviors exhibited therein. These correlations are not inherently obvious unless one has prior domain expertise, and exploration of these relationships can be tedious even with complex statistical analysis tools. Ultimately, public health officials seek simplified datasets providing the minimum factors necessary to illustrate a problematic scenario. Through a combination of specific health data sources and innovative text and data analyses, we created a visualization engine allowing public health officials to quickly define, assess, and address potential epidemics. Our innovations include unique text and data mining techniques for discovering relationships spanning multiple domains, and innovative interactive visualizations providing comprehensive knowledge transfer to the user.

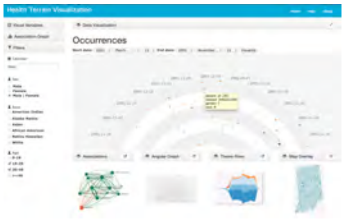
Introduction and Background

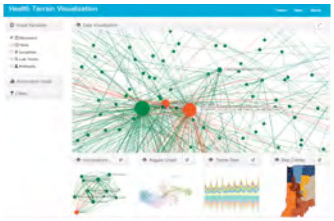
Public health researchers need access to relationships within communities to understand the details of an epidemic. Individuals can be connected to others by geographic proximity, shared community spaces, common treatments, or other shared personal traits. We found that diseases and mental behaviors also exhibit relationships over time, and these connections become more obvious in large numbers. Individual analysis of each patient, disease, and mental behavior against a given cohort could eventually assist in identifying the cause for an epidemic, but the work is tedious and requires extensive domain knowledge to be fruitful. An effective visualization system for public health officials should provide insight to these dimensions without requiring prior knowledge of dataset characteristics.

Our visualization interface provides multiple intelligent visualizations of regionalized health data, allowing users to see relationships at varying levels of granularity. We tailored the system's user interface for quick access to data filtering controls and snapshots of multiple alternative visualizations, surrounding the main interactive window that can assume the full screen on demand.

Unique data sources and analysis techniques coupled with interactive visualizations comprise the core of our innovations. The Notifiable Condition Detector (NCD) system at Regenstrief Institute informs public health officials by identifying lab tests indicating any reportable conditions, as defined by the state of Indiana. We use outcomes from the NCD to preselect patients of interest in our dataset. We then align these patients with discharge notes from hospitals in the Indiana Network for Patient Care, and process all of the free-form text data through text mining analysis to identify diseases and mental behaviors. Our system establishes correlations across domains, whether patient, disease, or mental behavior, and prepares this metadata for informing our unique visualizations: a navigable weighted graph connecting patients, diseases, and mental behaviors; a geospatial heat map showing population density of selected diseases; an angular graph depicting patient-disease relationships intensity over time; a Sankey diagram demonstrating the flow of patients from one disease to another over time; and a theme river used to show interactions among diseases and mental behaviors over time.







As of this writing, our system is still under development and will not be deployed publicly unless so determined by its funder. This project is supported by the Department of the Army, award number W81CWH-13-1-0020.